Analyse statistique des risques actuariels

Exercices et solutions

Analyse statistique des risques actuariels

Exercices et solutions

Marie-Pier Côté

École d'actuariat
Université Laval

Première édition préliminaire

Historique de publication
Janvier 2019 : Première édition préliminaire

Code source
Le code source LaTeX de la première édition de ce document est disponible en communiquant directement avec les auteurs.

Introduction

Ce document est une collection d'exercices distribués par l'auteure dans le cadre du cours ACT-2000 *Analyse statistique des risques actuariels* à l'École d'actuariat de l'Université Laval. Certains exercices sont le fruit de l'imagination des auteurs du recueil ou des versions précédentes, alors que plusieurs autres sont des adaptations d'exercices tirés des ouvrages cités dans la bibliographie.

C'est d'ailleurs afin de ne pas usurper de droits d'auteur que ce document est publié selon les termes du contrat Paternité-Partage des conditions initiales à l'identique 2.5 Canada de Creative Commons. Il s'agit donc d'un document «libre» que quiconque peut réutiliser et modifier à sa guise, à condition que le nouveau document soit publié avec le même contrat.

Le recueil d'exercices se veut un complément à un cours de statistique mathématique pour des étudiants de premier cycle universitaire. Les exercices sont divisés en six chapitres qui correspondent aux chapitres de notre cours. Le chapitre 1 porte sur les modèles statistiques de base et comprend la notion d'échantillon aléatoire, quelques rappels de probabilité, ainsi que la notion de statistique d'ordre. Il est suivi du chapitre 2 qui traite des distributions d'échantillonnage et présente les distributions liées à la loi normale, soit la loi t de Student, la loi du khi carré, et la distribution de Fisher–Snedecor.

Au chapitre 3, on aborde les diverses propriétés des estimateurs. Le chapitre 4 traite d'estimation ponctuelle par les méthodes classiques (maximum de vraisemblance, méthode des moments, etc.). Enfin, les notions étroitement liées d'estimation par intervalle et de test d'hypothèses font l'objet des chapitres 5 et 6.

Les réponses des exercices se trouvent à la fin de chacun des chapitres, alors que les solutions complètes sont regroupées à l'annexe B. De plus, on trouvera sur le site de cours une liste non exhaustive d'exercices proposés dans [6]. Des solutions de ces exercices sont offertes dans [7], ou encore sous forme de petits clips vidéo (*solutions clip*) disponibles dans le portail Libre de l'École d'actuariat à l'adresse

`http://libre.act.ulaval.ca`

L'annexe A contient des tables de quantiles des lois normale, khi carré, t et F. J'encourage le lecteur à utiliser le logiciel R [5] pour résoudre certains exercices.

Je remercie Vincent Mercier pour son aide dans la préparation de ce recueil. Je remercie aussi d'avance les lecteurs qui voudront bien me faire part de toute erreur ou omission dans les exercices ou leurs solutions.

Marie-Pier Côté <marie-pier.cote@act.ulaval.ca>
Québec, janvier 2019

Table des matières

1 Modèles statistiques de base

1.1 Est-ce que les énoncés suivants constituent des exemples d'échantillon aléatoire ? Justifiez votre réponse.

a) Le nombre annuel de cas de cancer du sein causant un décès au Québec entre 1970 et 2018.

b) Le résultat de 20 lancers de dés non truqués lors d'une partie d'un jeu de société.

1.2 Dans chacun des cas suivants, identifiez la loi de probabilité qui serait la plus appropriée selon le contexte de l'énoncé. Justifiez votre choix et spécifiez quels paramètres sont connus et lesquels sont inconnus.

a) La perte financière du propriétaire d'une maison rasée par les flammes.

b) Le nombre de sacs de grains de café devant être examinés avant de trouver 15 sacs contaminés.

c) Le nombre de sacs de grains de café contaminés parmi 15 sacs examinés.

d) La vitesse réelle d'un véhicule à un endroit spécifique sur l'autoroute.

e) La véritable résistance d'un câble utilisé dans un ordinateur.

1.3 La distribution de Weibull est fréquemment utilisée en assurance IARD pour la modélisation des montants de sinistres, entre autres. Sa fonction de répartition est

$$F(x) = 1 - e^{-\beta x^{\alpha}}, \quad x > 0, \alpha > 0, \beta > 0.$$

a) Déterminer la fonction de densité de probabilité de la Weibull.

b) Calculer l'espérance et la variance de la Weibull.

1.4 Obtenir l'expression de $E[X]$ et $\text{var}[X]$ quand la variable aléatoire X suit une distribution Beta(α,β). [Indices :

i. Pour tout $\alpha > 0$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.

ii. Si f est une densité, alors $\int_{\mathbb{R}} f(x)\mathrm{d}x = 1$.]

1.5 Calculer la fonction génératrice des moments M d'une distribution de Poisson avec paramètre $\lambda > 0$. Utilisez M pour montrer que l'espérance et la variance d'une variable aléatoire Poisson sont égales.

1.6 Une station service opère deux pompes à essence. Chacune d'entre elles peut produire jusqu'à 10 000 litres d'essence par mois. La quantité totale d'essence pompée à la station chaque mois est une variable aléatoire Y, mesurée en 10 000 litres, avec fonction de densité de probabilité donnée par

$$f(y) = \begin{cases} y, & 0 \leq y < 1, \\ 2 - y, & 1 \leq y < 2, \\ 0, & \text{ailleurs.} \end{cases}$$

a) Trouver la fonction de répartition F de Y.

b) Calculer la probabilité que la station service pompe entre 8 500 et 11 500 litres d'essence dans un mois.

c) Quel est le revenu mensuel espéré si la station service vend son essence au prix de 2,10 $ le litre ?

1.7 Soit X une variable aléatoire de moyenne μ et de variance σ^2. Déterminer la valeur de c qui minimise $E[(X - c)^2]$.

1.8 Soit $M_X(t)$ la fonction génératrice des moments de la variable aléatoire X.

a) Soit $Y = aX + b$, où a et b sont des constantes quelconques. Démontrer que

$$M_Y(t) = e^{bt} M_X(at).$$

b) Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et $Y = X_1 + \cdots + X_n$. Démontrer que

$$M_Y(t) = \prod_{j=1}^{n} M_{X_j}(t).$$

1.9 Soit $X_1, \ldots, X_n$ un échantillon aléatoire tiré d'une distribution avec densité

$$f(x) = \begin{cases} \frac{200}{x^3}, & x \geq 10 \\ 0, & \text{ailleurs} \end{cases}$$

Est-ce qu'il est possible d'utiliser la Loi faible des grands nombres pour cet exemple ? Justifiez.

1.10 Soit $X_1, \ldots, X_n$ un échantillon aléatoire tiré d'une distribution avec densité

$$f(x) = 4(1 - x)^3, \quad \text{pour } 0 \leq x \leq 1.$$

Montrer que $\bar{X}_n$ converge en probabilité vers une contante et trouver cette constante.

1.11 Soit $X_1, \ldots, X_n$ un échantillon aléatoire tiré d'une loi avec fonction de répartition $F_X(\cdot)$ et $X_{(1)} \leq \cdots \leq X_{(n)}$ les statistiques d'ordre correspondantes. Trouver la fonction de répartition de $X_{(1)} = \min(X_1, \ldots X_n)$.

1.12 Soient $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq X_{(4)}$ les statistiques d'ordre d'un échantillon aléatoire de taille 4 issu d'une distribution avec fonction de densité de probabilité $f(x) = e^{-x}$, $0 < x < \infty$. Calculer $\Pr[X_{(4)} > 3]$.

1.13 Soit X_1, X_2, X_3 un échantillon aléatoire issu d'une loi bêta de paramètres $\alpha = 2$ et $\beta = 1$. Calculer la probabilité que la plus petite valeur de l'échantillon soit supérieure à la médiane (théorique) de la distribution.

1.14 Soit X une variable aléatoire discrète avec fonction de masse de probabilité $\Pr[X = x] = 1/6$, $x = 1,2,3,4,5,6$. Démontrer que la fonction de masse de probabilité du minimum d'un échantillon aléatoire de taille 5 issu de cette distribution est

$$\Pr\big[X_{(1)} = x\big] = \left(\frac{7-x}{6}\right)^5 - \left(\frac{6-x}{6}\right)^5, \quad x = 1,2,3,4,5,6.$$

1.15 Soient $X_{(1)} \leq \cdots \leq X_{(n)}$ les statistiques d'ordre d'un échantillon aléatoire tiré d'une loi de Weibull, dont la fonction de répartition est $F_X(x) = 1 - e^{-\beta x^\alpha}$. Calculer la fonction de répartition, la fonction de densité et l'espérance de $X_{(1)}$.

1.16 Soit un échantillon aléatoire $X_1,\ldots,X_n$ tiré d'une distribution F avec densité $f(x)$, pour $x \in \mathbb{R}$. Trouver la fonction de densité conjointe de $X_{(1)} = \min(X_1,\ldots X_n)$ et de $X_{(n)} = \max(X_1,\ldots X_n)$, $f_{X_{(1)},X_{(n)}}(x,y)$.

1.17 Soit un échantillon aléatoire $X_1,\ldots,X_n$. L'*étendue* (ou *dispersion*) empirique est

$$R = X_{(n)} - X_{(1)},$$

et la *mi-étendue* (ou *mi-dispersion*) empirique est

$$T = \frac{X_{(1)} + X_{(n)}}{2}.$$

En utilisant le fait que la densité conjointe de $X_{(1)}$ et de $X_{(n)}$ est donnée, pour $x < y$, par

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)(F_X(y) - F_X(x))^{n-2} f_X(x) f_X(y),$$

trouver la distribution conjointe de (R,T).

1.18 Si $X \sim \mathcal{U}(0,1)$, montrer que la densité de l'étendue est $f_R(r) = n(n-1)r^{n-2}(1-r)$, pour $0 < r < 1$.

1.19 Calculer la probabilité que l'étendue d'un échantillon aléatoire de taille 4 issu d'une loi uniforme sur l'intervalle $(0,1)$ soit inférieure à $1/2$.

1.20 Si un échantillon de taille 2 est tiré d'une loi bêta avec paramètres $\alpha = 1$ et $\beta = 2$, quelle est la probabilité que l'une des deux valeurs de l'échantillon soit au moins deux fois plus grande que l'autre ? (*Astuce* : intégrer la densité conjointe des deux valeurs de l'échantillon au-dessus de la surface correspondant à la probabilité recherchée.)

1.21 Soit $X \sim \mathcal{U}(0,1)$. Calculer l'espérance de la mi-étendue $T = (X_{(1)} + X_{(n)})/2$ d'un échantillon de taille n issu de cette distribution.

1.22 Soit $X_1,\ldots,X_n$ un échantillon aléatoire d'une loi uniforme sur l'intervalle $(0,1)$.

a) Calculer la moyenne et la variance de $R = X_{(n)} - X_{(1)}$.

b) Calculer la moyenne et la variance de $T = (X_{(1)} + X_{(n)})/2$.

Réponses

1.1 a) non b) oui

1.3 a) $f(x) = \alpha\beta x^{\alpha-1}e^{-\beta x^\alpha}$ b) $E[X] = \beta^{-1/\alpha}\Gamma(1+1/\alpha)$, $\text{var}[X] = \beta^{-2/\alpha}(\Gamma(1+2/\alpha) - \Gamma(1+1/\alpha)^2)$

1.4 $\text{E}(X) = \alpha/(\alpha+\beta)$ et $\text{var}[X] = (\alpha\beta)/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$

1.5 $M_X(t) = \exp\{\lambda(e^t - 1)\}$ pour tout $t \in \mathbb{R}$.

1.6 b) 0.2775 c) 21 000

1.7 $c = \mu$

1.9 Non.

1.10 $1/5$

1.11 $F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n$

1.12 $1 - (1 - e^{-3})^4$

1.13 $1/8$

1.15 $X_{(1)} \sim \text{Weibull}(\alpha, n\beta)$.

1.16 $f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)(F(y) - F(x))^{n-2}f(x)f(y), \quad x < y$

1.17 $f_{R,T}(r,t) = n(n-1)\{F_X(t + r/2) - F_X(t - r/2)\}^{n-2}f_X(t + r/2)f_X(t - r/2)$, pour $r > 0$ et $-\infty < t < \infty$.

1.19 $5/16$

1.20 $7/12$

1.21 $1/2$.

1.22 a) $E[R] = (n-1)/(n+1)$ et $\text{var}[R] = (2n-2)/[(n+1)^2(n+2)]$.
b) $E[T] = 1/2$ et $\text{var}[T] = 1/[2(n+1)(n+2)]$.

2 Distributions d'échantillonnage

2.1 Soit $\bar{X}_5$ la moyenne d'un échantillon de taille 5 tiré d'une distribution normale avec moyenne 10 et variance 125. Trouver la constante c telle que $\Pr[\bar{X}_5 < c] = 0{,}90$.

2.2 Si $\bar{X}_n$ est la moyenne d'un échantillon de taille n tiré d'une distribution normale de moyenne μ et de variance 100, trouver la valeur de n telle que $\Pr\left[\mu - 5 < \bar{X}_n < \mu + 5\right] = 0{,}954$.

2.3 Soit $X_1, \ldots, X_{25}$ un échantillon aléatoire issu d'une distribution $\mathcal{N}(0,16)$ et $Y_1, \ldots, Y_{25}$ un échantillon aléatoire issu d'une distribution $\mathcal{N}(1,9)$. Les deux échantillons sont indépendants. Soient $\bar{X}_{25}$ et $\bar{Y}_{25}$ les moyennes des deux échantillons. Calculer la probabilité $\Pr[\bar{X}_{25} > \bar{Y}_{25}]$.

2.4 Soit $Y_1, \ldots, Y_8$ un échantillon aléatoire issu d'une distribution $\mathcal{N}(0,1)$ et

$$\bar{Y} = (Y_1 + \cdots + Y_7)/7.$$

Quelle est la distribution des statistiques suivantes ? Justifiez vos réponses.

a) $W = \sum_{i=1}^{7} Y_i^2$

b) $U = \sum_{i=1}^{7} (Y_i - \bar{Y})^2$

c) $Y_8^2 + U$

d) $\sqrt{7} Y_8 / \sqrt{W}$

e) $\sqrt{6} Y_8 / \sqrt{U}$

f) $3(7\bar{Y}^2 + Y_8^2)/U$

2.5 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une loi normale de moyenne μ et variance σ^2. Trouver la moyenne et la variance de la statistique

$$S_n^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

2.6 Soit S_6^2 la variance d'un échantillon de taille 6 d'une distribution normale de moyenne μ et de variance 10. Calculer $\Pr[2{,}30 < S_6^2 < 22{,}2]$.

2.7 Trouvez la fonction de densité de probabilité de S_n^2, la variance échantillonale d'un échantillon de taille n d'une distribution $\mathcal{N}(0,\sigma^2)$. Est-ce que la distribution échantillonale de S_n^2 fait partie d'une famille de distributions connues ? Utilisez ce résultat

pour trouver la moyenne et la variance de S_n^2. [*Truc : Considérez* $(n-1)S_n^2/\sigma^2 \sim \chi^2_{(n-1)}$ *et utilisez la méthode de la fonction de répartition pour obtenir la densité de la variable aléatoire transformée.*]

2.8 Soit $Z \sim \mathcal{N}(0,1)$ avec fonction de densité de probabilité

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, \quad -\infty < z < \infty,$$

et fonction de répartition

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\,dx.$$

Exprimer les fonctions de densité et de répartition de $X = \mu + \sigma Z$ en fonction de $\phi(\cdot)$ et $\Phi(\cdot)$.

2.9 Soit la variable aléatoire $Z \sim \mathcal{N}(0,1)$. Démontrer que $Z^2 \sim \chi^2(1)$ avec la technique de la fonction génératrice des moments. (*Note* : il faut intégrer pour trouver la fonction génératrice des moments de Z^2.)

2.10 a) Soit $X \sim \mathcal{N}(0,\sigma^2)$. Trouver la distribution de $Y = X^2$.

b) Soient X_1 et X_2 deux variables aléatoires indépendantes chacune distribuée selon une loi normale centrée réduite. Trouver la distribution de

$$Y = \frac{(X_1 - X_2)^2}{2}.$$

2.11 Soit T une variable aléatoire distribuée selon une loi t avec 10 degrés de liberté.

a) Trouver $\Pr[|T| > 2{,}228]$ à l'aide d'une table de la loi t.

b) Répéter la partie a) à l'aide R. La fonction `pt(x, n)` donne la valeur de la fonction de répartition en `x` d'une loi t avec `n` degrés de liberté.

2.12 Soit T une variable aléatoire distribuée selon une loi t avec 14 degrés de liberté.

a) Trouver la valeur de b tel que $\Pr[-b < T < b] = 0{,}90$ à l'aide d'une table de la loi t.

b) Répéter la partie a) à l'aide R. La fonction `qt(p, n)` retourne le $\mathtt{p}^{\mathrm{e}}$ quantile d'une loi t avec `n` degrés de liberté, c'est-à-dire la valeur de x où la fonction de répartition vaut `p`.

2.13 Soit $U \sim \chi^2(r_1)$ et $V \sim \chi^2(r_2)$, deux variables aléatoires indépendantes.

a) Démontrer que la densité de

$$F = \frac{U/r_1}{V/r_2}$$

est

$$f(x) = \frac{\Gamma((r_1+r_2)/2)(r_1/r_2)^{r_1/2}x^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1x/r_2)^{(r_1+r_2)/2}}.$$

b) Calculer $E[F]$.

c) Calculer $\text{var}[F]$.

2.14 Soit F une variable aléatoire distribuée selon une loi F avec ν_1 et ν_2 degrés de liberté (dans l'ordre). Démontrer que $1/F$ est aussi distribuée selon une loi F, mais avec ν_2 et ν_1 degrés de liberté.

2.15 Si F a une distribution F avec paramètres $\nu_1 = 5$ et $\nu_2 = 10$, trouver a et b de sorte que $\Pr[F \leq a] = 0{,}05$ et $\Pr[F \leq b] = 0{,}95$. Les quantiles de la loi F peuvent être trouvés soit dans une table, soit à l'aide de la fonction `qf(x, v1, v2)` de R. (*Astuce* : en travaillant avec une table, utiliser le fait que $\Pr[F \leq a] = \Pr[F^{-1} \geq a^{-1}] = 1 - \Pr[F^{-1} \leq a^{-1}]$.)

2.16 Soit $T = W/\sqrt{V/r}$, où W et V sont des variables aléatoires indépendantes avec une distribution, respectivement, normale centrée réduite et khi carré avec r degrés de liberté. Démontrer que la distribution de T^2 est F avec 1 et r degrés de liberté.

2.17 Démontrer à l'aide du Théorème central limite que la distribution gamma avec paramètre de forme α entier et paramètre d'échelle β tend vers la distribution normale avec moyenne $\alpha\beta$ et variance $\alpha\beta^2$ lorsque α tend vers l'infini. (*Astuce* : définir $Y = X_1 + \cdots + X_\alpha$ où $X_i \sim \text{Exponentielle}(\beta)$ et trouver la distribution asymptotique de Y.)

2.18 Soit $\bar{X}_{100}$ la moyenne d'un échantillon aléatoire de taille 100 tiré d'une loi $\chi^2(50)$.

a) Trouver la distribution exacte de $\bar{X}_{100}$.

b) Calculer à l'aide d'un logiciel statistique la valeur exacte de $\Pr[49 < \bar{X}_{100} < 51]$.

c) Calculer une valeur approximative de la probabilité en b).

2.19 Soit $\bar{X}$ la moyenne d'un échantillon de taille 128 d'une loi Gamma(2,4). Trouver une approximation pour $\Pr[7 < \bar{X} < 9]$.

2.20 Trouver une valeur approximative de la probabilité que la moyenne d'un échantillon de taille 15 d'une loi avec densité $f(x) = 3x^2$, $0 < x < 1$, soit entre $3/5$ et $4/5$.

2.21 On suppose que $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_n$ sont des échantillons aléatoires indépendants de populations avec moyennes μ_1 et μ_2 et variances $\sigma_1^2 > 0$ et $\sigma_2^2 > 0$, respectivement. Démontrer que, quand $n \to \infty$,

$$U_n = \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}}$$

est asymptotiquement $\mathcal{N}(0,1)$.

2.22 Une machine d'embouteillage peut être réglée de sorte qu'elle remplisse en moyenne les bouteilles de μ onces de liquide par bouteille. Il a été observé que la quantité de liquide distribuée par la machine suit une loi normale avec $\sigma = 2{,}5$ onces.

a) Si $n = 9$ bouteilles sont sélectionnées aléatoirement à la sortie de la machine, quelle est la probabilité que la moyenne échantillonnale diffère de μ d'au plus 0,2 once ?

b) Trouver la probabilité que la moyenne échantillonale diffère de μ d'au plus 0,2 once lorsque la taille de l'échantillon est $n = 25, 36$, et 64. Que remarquez-vous lorsque n augmente ? Pouvez-vous expliquer cette remarque ?

c) Quelle est la taille d'échantillon requise pour s'assurer que la probabilité que la moyenne échantillonnale diffère de μ d'au plus 0,2 once est d'au moins 95% ?

d) Comment la probabilité obtenue en a) change-t-elle lorsque σ est inconnu et que la variance échantillonnale est égale à $s_n^2 = 5,5$?

2.23 L'*Agence de Protection de l'Environnement* est responsable d'établir les critères pour la quantité autorisée de certains produits chimiques en eau douce. Une mesure commune de la toxicité pour les polluants est la concentration d'un produit chimique causant la mort de la moitié d'une population animale donnée dans un laps de temps connu. Cette mesure est appelée CL50 (concentration létale médiane). Dans plusieurs études, les valeurs observées du logarithme naturel de l'indicateur CL50 sont normalement distribuées. L'analyse est donc basée sur les données de ln(CL50).

Soit S_n^2 la variance échantillonnale d'un échantillon de $n = 10$ valeurs de ln(CL50) pour le cuivre, et S_m^2 la variance échantillonnale d'un échantillon de $m = 6$ valeurs de ln(CL50) pour le plomb, tous les deux utilisant la même espèce de poisson. La variance de la population des mesures sur le cuivre est supposée être le double de la variance de la population des mesures sur le plomb. Supposer que S_n^2 est indépendante de S_m^2.

a) Expliquer comment il est possible, en référant à une table statistique appropriée, de trouver les nombres a et b tels que

$$\Pr\left(\frac{S_n^2}{S_m^2} \leq b\right) = 0,95, \quad \Pr\left(\frac{S_n^2}{S_m^2} \geq a\right) = 0,95.$$

Astuce : Observer que $\Pr[U_1/U_2 \leq k] = \Pr[U_2/U_1 \geq 1/k]$.

b) Si a et b sont les mêmes qu'en a), calculer

$$\Pr\left(a \leq \frac{S_n^2}{S_m^2} \leq b\right).$$

Réponses

2.1 16,41

2.2 16

2.3 0,159

2.5 $E[S_n^2] = \sigma^2$, $\text{var}[S_n^2] = 2\sigma^4/(n-1)$

2.6 0,90

2.8 $f_X(x) = \sigma^{-1}\phi((x-\mu)/\sigma)$, $F_X(x) = \Phi((x-\mu)/\sigma)$

2.10 a) Gamma$(1/2, 2\sigma^2)$ b) $\chi^2(1)$

2.11 0,05

2.12 1,761

2.13 b) $r_2/(r_2-2)$

c) $2[r_2^2(r_2+r_1-2)]/[r_1(r_2-2)^2(r_2-4)]$

2.15 $a = 0{,}211$ et $b = 3{,}33$

2.18 a) Gamma$(2\,500, 1/50)$ b) 0,6827218 c) 0,682

2.19 0,954

2.20 0,840

2.22 a) 0,1896 b) 0,3108 ; 0,3688 ; 0,4778 c) 601 d) 0,1955

2.23 a) 0,5744 b) 0,9

3 Estimation

3.1 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une distribution avec moyenne μ et variance σ^2. Démontrer que $n^{-1}\sum_{i=1}^{n}(X_i - \mu)^2$ est un estimateur sans biais de σ^2.

3.2 Si $X_1, \ldots, X_n$ est un échantillon aléatoire d'une distribution avec moyenne μ, quelle condition doit-on imposer sur les constantes $a_1, \ldots, a_n$ pour que

$$a_1 X_1 + \cdots + a_n X_n$$

soit un estimateur sans biais de μ ?

3.3 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une distribution avec moyenne μ et variance σ^2.

a) Démontrer que $\bar{X}_n^2$ est un estimateur biaisé de μ^2 et calculer son biais.

b) Démontrer que $\bar{X}_n^2$ est un estimateur asymptotiquement sans biais de μ^2.

3.4 Soient $X_{(1)} < X_{(2)} < X_{(3)}$ les statistiques d'ordre d'un échantillon aléatoire de taille 3 tiré d'une distribution uniforme avec fonction de densité

$$f(x) = \theta^{-1}, \quad 0 < x < \theta, \quad \theta > 0.$$

Démontrer que $4X_{(1)}$ et $2X_{(2)}$ sont tous deux des estimateurs sans biais de θ. Trouver la variance de chacun de ces estimateurs.

3.5 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une distribution uniforme sur l'intervalle $(0, \theta)$.

a) Développer un estimateur sans biais de θ basé sur $\max(X_1, \ldots, X_n)$.

b) Répéter la partie a), mais cette fois à partir de $\min(X_1, \ldots, X_n)$.

3.6 Soit $X \sim \text{Binomiale}(n, p)$. Démontrer que, malgré que X/n soit un estimateur sans biais de p,

$$n\left(\frac{X}{n}\right)\left(1 - \frac{X}{n}\right)$$

est un estimateur biaisé de la variance de X. Calculer le biais de l'estimateur.

3.7 Démontrer, à partir de la définition, que $X_{(1)} = \min(X_1, \ldots, X_n)$ est un estimateur convergent du paramètre θ d'une loi uniforme sur l'intervalle $(\theta, \theta + 1)$.

3.8 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une loi exponentielle de moyenne θ. Démontrer que $\bar{X}$ est un estimateur convergent de θ.

3.9 Soit $X_1,\ldots,X_n$ un échantillon aléatoire de taille $n \geq 3$ d'une population avec moyenne μ et variance $\sigma^2 > 0$. On considère les trois estimateurs suivants pour μ :

$$\hat{\mu}_1 = \frac{X_1 + X_2}{2}, \qquad \hat{\mu}_2 = \frac{X_1}{4} + \frac{X_2 + \cdots + X_{n-1}}{2(n-2)} + \frac{X_n}{4}, \quad \text{et} \quad \hat{\mu}_3 = \sum_{j=1}^{n} \frac{X_j}{n}.$$

a) Montrer que ces estimateurs sont sans biais.

b) Trouver l'efficacité de $\hat{\mu}_3$ par rapport à $\hat{\mu}_2$ et $\hat{\mu}_1$, respectivement.

c) Quel estimateur est préférable et pourquoi ?

3.10 Soit $X_1,\ldots,X_n$ un échantillon aléatoire avec fonction de répartition

$$F(x) = \begin{cases} 0, & x < \beta \\ 1 - (\beta/x)^{\alpha}, & x \geq \beta, \end{cases}$$

où α,$\beta > 0$. La fonction de densité de probabilité correspondante est

$$f(x) = \begin{cases} \alpha\beta^{\alpha} x^{-(\alpha+1)}, & x \geq \beta, \\ 0, & \text{ailleurs.} \end{cases}$$

a) Trouver la fonction de répartition de $\hat{\beta} = \min(X_1,\ldots,X_n)$.

b) Montrer que $\hat{\beta}$ est un estimateur convergent de β.

c) Calculer le biais et l'erreur quadratique moyenne de l'estimateur $\hat{\beta}$. Est-ce que cet estimateur est sans biais ou asymptotiquement sans biais ?

d) Si α est connu, montrer que $\min(X_1,\ldots,X_n)$ est une statistique exhaustive pour β.

e) Si β est connu, montrer que $X_1 \times \cdots \times X_n$ est une statistique exhaustive pour α.

f) Trouver une paire de statistiques conjointement exhaustives dans le cas où α et β sont inconnus.

3.11 Soit X_1 une observation d'une loi normale avec moyenne 0 et variance σ^2, $\sigma > 0$. Démontrer que $|X_1|$ est une statistique exhaustive pour σ^2.

3.12 Trouver une statistique exhaustive pour le paramètre θ de la loi uniforme sur l'intervalle $(-\theta,\theta)$.

3.13 Démontrer que la somme des éléments d'un échantillon aléatoire issu d'une loi de Poisson est une statistique exhaustive pour le paramètre de cette loi.

3.14 Soit $X_1,\ldots,X_n$ un échantillon aléatoire tiré d'une loi de Poisson avec paramètre λ inconnu. On sait que $T(X_1,\ldots,X_n) = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour λ. On estime λ par $\tilde{\lambda} = X_1$.

a) Montrer que $\tilde{\lambda}$ est un estimateur sans biais de λ.

b) Utiliser le théorème de Rao–Blackwell pour trouver un estimateur λ^* à partir de $\tilde{\lambda}$.

3.15 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une loi géométrique avec fonction de masse de probabilité

$$\Pr[X = x] = \theta(1-\theta)^x, \quad x = 0, 1, \ldots$$

Démontrer que $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ est une statistique exhaustive pour θ.

3.16 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une loi de Poisson avec moyenne λ.

a) Démontrer que $T = \sum_{i=1}^n X_i$ est exhaustive minimale pour λ.

b) Est-ce que $\bar{X}_n = T/n$ est un estimateur sans biais de variance minimale (MVUE) pour λ ? Expliquer.

c) L'*inégalité de Cramér–Rao–Fréchet* indique que si $\hat{\lambda}_n$ est un estimateur sans biais de λ, alors

$$\operatorname{var}(\hat{\lambda}_n) \geq \left\{ n \operatorname{E}\left[-\frac{\partial^2}{\partial \lambda^2} \ln f(X;\lambda) \right] \right\}^{-1}.$$

Un estimateur avec variance égale à la borne inférieure est dit être *efficace*. Montrer que $\bar{X}_n$ est en effet un estimateur efficace pour λ.

3.17 Démontrer que, sous les hypothèses appropriées,

$$E\left[\left(\frac{\partial}{\partial \theta} \ln f(X;\theta) \right)^2 \right] = -E\left[\frac{\partial^2}{\partial \theta^2} \ln f(X;\theta) \right].$$

Pour ce faire, dériver par rapport à θ l'identité

$$\int_{-\infty}^{\infty} f(x;\theta)\, dx = 1$$

afin d'obtenir

$$\int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \theta} \ln f(x;\theta) \right) f(x;\theta)\, dx = 0,$$

puis dériver de nouveau par rapport à θ.

3.18 Démontrer que la moyenne arithmétique est un estimateur sans biais à variance minimale du paramètre λ d'une loi de Poisson.

3.19 Démontrer que la proportion de succès X/n est un estimateur sans biais à variance minimale de la probabilité de succès θ d'une distribution Binomiale. (*Astuce* : considérer X/n comme la moyenne d'un échantillon aléatoire d'une distribution de Bernoulli.)

3.20 Supposons que $\bar{X}_1$ est la moyenne d'un échantillon aléatoire de taille n d'une population normale avec moyenne μ et variance σ_1^2, que $\bar{X}_2$ est la moyenne d'un échantillon aléatoire de taille n d'une population normale avec moyenne μ et variance σ_2^2 et que les deux échantillons aléatoires sont indépendants.

a) Démontrer que $\omega \bar{X}_1 + (1-\omega)\bar{X}_2$, $0 \leq \omega \leq 1$, est un estimateur sans biais de μ.

b) Démontrer que la variance de $\omega \bar{X}_1 + (1-\omega)\bar{X}_2$ est minimale lorsque

$$\omega = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

c) Calculer l'efficacité relative de l'estimateur en a) avec $\omega = \frac{1}{2}$ à celle de l'estimateur à variance minimale trouvé en b).

Réponses

3.2 $\sum_{i=1}^{n} a_i = 1$

3.3 a) σ^2/n

3.4 $\text{var}[4X_{(1)}] = 3\theta^2/5$, $\text{var}[2X_{(2)}] = \theta^2/5$.

3.5 a) $(n+1)X_{(n)}/n$ b) $(n+1)X_{(1)}$

3.6 $-p(1-p)$

3.9 b) $\frac{n^2}{8(n-2)}$ et $\frac{n}{2}$ c) $\hat{\mu}_3$

3.12 $\max_{i=1,\ldots,n}(|X_i|)$, ou $(X_{(1)}, X_{(n)})$

3.20 c) $(\sigma_1^2 + \sigma_2^2)^2/(4\sigma_1^2\sigma_2^2)$

4 Ajustement de modèles

4.1 Soit $X_1, \ldots, X_n$ un échantillon aléatoire issu des distributions ci-dessous. Dans chaque cas, trouver l'estimateur du paramètre θ à l'aide de la méthode des moments.

a) $f(x;\theta) = \theta^x e^{-\theta}/x!$, $x = 0,1,\ldots$, $\theta > 0$.

b) $f(x;\theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$.

c) $f(x;\theta) = \theta^{-1} e^{-x/\theta}$, $\theta > 0$.

d) $f(x;\theta) = e^{-|x-\theta|}/2$, $-\infty < x < \infty$, $-\infty < \theta < \infty$.

e) $f(x;\theta) = e^{-(x-\theta)}$, $x \geq \theta$, $\theta > 0$.

4.2 Si $Z \sim \mathcal{N}(\mu,\sigma^2)$ et $Y = e^Z$, alors Y a une distribution log-normale avec paramètres μ et σ^2. La fonction de répartition est

$$F(y) = \Phi\left(\frac{\ln(y) - \mu}{\sigma}\right), \quad y > 0,$$

où Φ est la fonction de répartition de la loi normale standard.

a) Utiliser la fonction génératrice des moments de la loi normale pour montrer que les deux premiers moments de la loi log-normale sont

$$\mathrm{E}[Y] = e^{\mu+\sigma^2/2} \quad \text{et } \mathrm{E}[Y^2] = e^{2\mu+2\sigma^2}.$$

b) Utiliser la méthode des moments pour construire un estimateur de μ et σ^2 si $Y_1, \ldots, Y_n$ forment un échantillon aléatoire d'une loi log-normale.

c) Montrer que les estimateurs trouvés en (b) sont convergents.

4.3 Considérer la distribution géométrique avec fonction de masse de probabilité

$$\Pr[X = x] = \theta(1-\theta)^x, \quad x = 0,1,\ldots.$$

On a obtenu l'échantillon aléatoire suivant de cette distribution :

5 7 4 11 0 9 1 1 3 2 1 0 6 0 1 1 1 9 2 0.

Utiliser la méthode des moments pour obtenir une estimation ponctuelle du paramètre θ.

4.4 On a un dé régulier avec θ faces numérotées de 1 à θ, où θ est un entier inconnu. Soit $X_1, \ldots, X_n$ un échantillon aléatoire composé de n lancers de dés indépendants.

a) Déterminer l'estimateur des moments de θ.

b) Calculer l'estimateur des moments de θ si $n = 4$ et $x_1 = x_2 = x_3 = 3$ et $x_4 = 12$. Interpréter le résultat.

4.5 Soit $X_1, \ldots, X_n$ un échantillon aléatoire tiré d'une distribution log-normale. Les 20e et 80e quantiles empiriques sont respectivement de $\hat{\pi}_{0,20} = 18{,}25$ et $\hat{\pi}_{0,80} = 35{,}8$. Estimer les paramètres de la distribution en utilisant la méthode des quantiles, puis utiliser ces estimations pour estimer la probabilité d'observer une valeur excédant 30.

4.6 Un échantillon aléatoire de réclamations est tiré d'une distribution log-logistique avec fonction de répartition

$$F(x) = \frac{(x/\theta)^\tau}{1 + (x/\theta)^\tau}, \quad x > 0, \theta > 0, \tau > 0.$$

Dans l'échantillon, 80% des réclamations excèdent 100 et 20% des réclamations excèdent 400. Estimer les paramètres θ et τ par la méthode des quantiles.

4.7 Les 20 pertes suivantes (en millions de dollars) ont été enregistrées sur une période d'un an :

1	1	1	1	1	2	2	3	3	4
6	6	8	10	13	14	15	18	22	25

Déterminer le 75e quantile empirique par la méthode des quantiles empiriques lissés.

4.8 Trouver l'estimateur du maximum de vraisemblance du paramètre θ de chacune des distributions de l'exercice 4.1.

4.9 Soit $X_1, \ldots, X_n$ un échantillon aléatoire de la distribution exponentielle translatée avec fonction de répartition

$$F(x; \mu, \lambda) = 1 - e^{-\lambda(x-\mu)}$$

et densité

$$f(x; \mu, \lambda) = \lambda e^{-\lambda(x-\mu)}, \quad x \geq \mu,$$

où $-\infty < \mu < \infty$ et $\lambda > 0$.

a) Démontrer que la distribution exponentielle translatée est obtenue par la transformation $X = Z + \mu$, où $Z \sim$ Exponentielle(λ) et $E[Z] = \frac{1}{\lambda}$.

b) Calculer l'espérance et la variance de cette distribution.

c) Calculer les estimateurs du maximum de vraisemblance des paramètres μ et λ.

d) Simuler 100 observations d'une loi Exponentielle translatée de paramètres $\mu = 1000$ et $\lambda = 0{,}001$ à l'aide de la fonction `rexp` de R et de la transformation en a). Calculer des estimations ponctuelles de μ et λ pour l'échantillon ainsi obtenu. Ces estimations sont-elles proches des vraies valeurs des échantillons ? Répéter l'expérience plusieurs fois au besoin.

4.10 Soient $X_{(1)} < \cdots < X_{(n)}$ les statistiques d'ordre d'un échantillon aléatoire tiré d'une distribution uniforme sur l'intervalle $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, $-\infty < \theta < \infty$. Démontrer que toute statistique $T(X_1, \ldots, X_n)$ satisfaisant l'inégalité

$$X_{(n)} - \frac{1}{2} \leq T(X_1, \ldots, X_n) \leq X_{(1)} + \frac{1}{2}$$

est un estimateur du maximum de vraisemblance de θ. Ceci est un exemple où l'estimateur du maximum de vraisemblance n'est pas unique.

4.11 Soit $X_1, \ldots, X_n$ un échantillon aléatoire issu de la distribution inverse gaussienne, dont la densité est

$$f(x; \mu, \lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x > 0.$$

Calculer les estimateurs du maximum de vraisemblance de μ et λ.

4.12 Soit $X_1, \ldots, X_n$ un échantillon aléatoire issu d'une distribution Pareto type II tel que pour tout $i \in \{1, \ldots, n\}$ et $x > 0$,

$$F(x) = 1 - \left(\frac{\theta}{x+\theta}\right)^{\alpha} \quad \text{et} \quad f(x) = \frac{\alpha\theta^{\alpha}}{(x+\theta)^{\alpha+1}},$$

avec paramètres inconnus $\alpha > 0$ et $\theta > 0$.

a) Trouver $\mathrm{E}[X_1]$ et $\mathrm{var}(X_1)$.

b) On suppose $\alpha > 2$. Construire des estimateurs pour les paramètres α et θ en utilisant la méthode des moments.

c) Écrire la vraisemblance, la log-vraisemblance et les deux équations qui doivent être résolues pour trouver les estimateurs du maximum de vraisemblance pour α et θ.

4.13 Soit $X_1, \ldots, X_n$ un échantillon aléatoire issu d'une loi uniforme sur l'intervalle (a, b) où a et b sont des constantes inconnues. Calculer l'estimateur du maximum de vraisemblance de a et b.

4.14 On suppose que $X_1, \ldots, X_n$ forment un échantillon aléatoire d'une loi uniforme sur l'intervalle $(0, 2\theta + 1)$ pour un paramètre inconnu $\theta > -1/2$.

a) Calculer l'estimateur du maximum de vraisemblance pour θ.

b) Calculer l'EMV de $\mathrm{var}(X)$, où X suit une distribution $\mathcal{U}(0, 2\theta + 1)$.

4.15 Soit $X_1, \ldots, X_n$ un échantillon aléatoire issu d'une distribution dont la loi de probabilité est

$$\Pr[X = x] = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1, \quad 0 \leq \theta \leq \frac{1}{2}.$$

a) Calculer les estimateurs du maximum de vraisemblance et des moments de θ.

b) Calculer l'erreur quadratique moyenne pour les estimateurs développés en a).

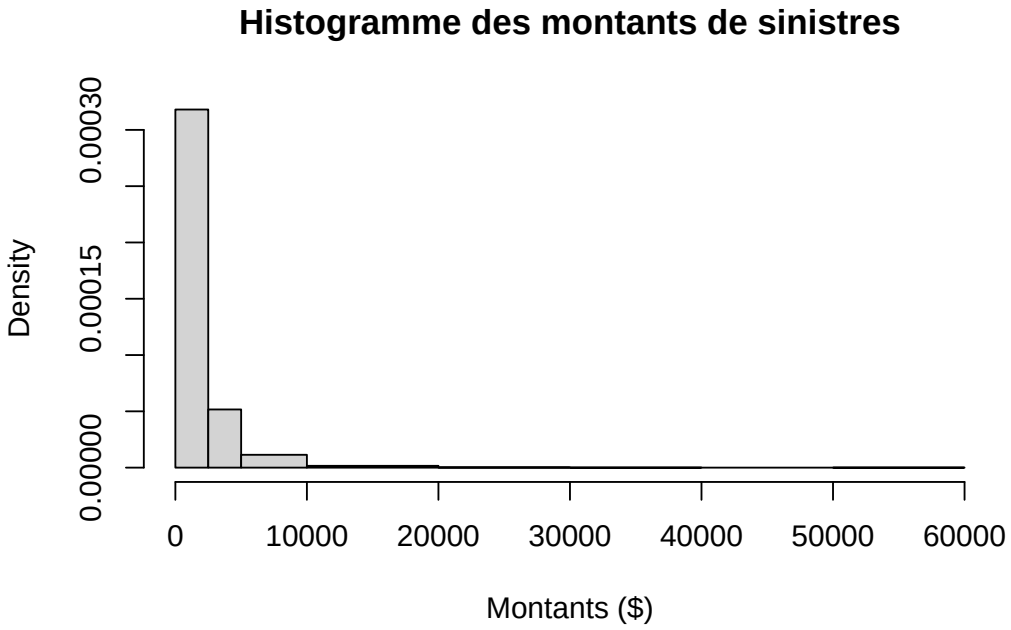
c) Lequel des estimateurs obtenus en a) est le meilleur ? Justifier.

4.16 Un assureur IARD américain souhaite modéliser les montants des sinistres pour une assurance automobile privée. Les données disponibles contiennent $n = 6773$ réclamations, et $X_1, \ldots, X_n$ représentent les montants en dollars US. Les réclamations sont considérées indépendantes. Un sommaire et un histogramme des réclamations sont fournis ci-dessous. On a également

$$\frac{1}{n}\sum_{i=1}^{n} x_i = 1\,853 \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 10\,438\,832.$$

```
summary(PAID)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.5   523.7  1001.7  1853.0  2137.4 60000.0
```



a) En utilisant les informations fournies, argumenter qu'une loi normale ne devrait pas être utilisée pour modéliser les montants de réclamations. Proposer (au moins) une distribution qui semble appropriée et justifier.

b) On suppose que les montants de réclamations suivent une loi log-normale définie à l'exercice 4.2. Trouver les estimateurs des moments des paramètres μ et σ^2.

c) On suppose que les montants de réclamations sont distribués selon une loi Pareto type II définie à l'exercice 4.12. Trouver les estimateurs des paramètres α et θ par la méthode des moments.

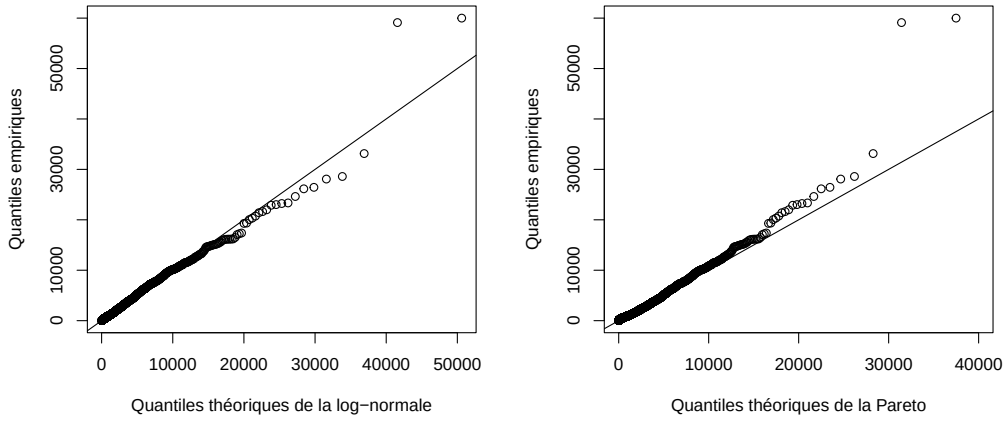
d) Les estimateurs du maximum de vraisemblance des distributions log-normale et Pareto peuvent être calculés par optimisation numérique en utilisant les estimateurs des moments trouvés en b) et c) comme valeurs de départ. Les résultats sont les suivants :

Log-normale : Les estimateurs sont $\hat{\mu}_{MV} = 6{,}95561$ et $\hat{\sigma}^2_{MV} = 1{,}14698$ et la valeur de log-vraisemblance est $-57\,185{,}11$.

Pareto : Les estimateurs sont $\hat{\alpha}_{MV} = 4{,}71364$ et $\hat{\theta}_{MV} = 6\,819{,}891$ et la valeur de log-vraisemblance est $-57\,500{,}12$.

Calculer l'AIC pour chacune des distributions. Quelle distribution semble être préférable selon ce critère ?

e) Les diagrammes quantile-quantile de chacun des modèles sont présentés ci-dessous. Commenter sur l'ajustement des modèles. Quel modèle recommandez-vous à l'assureur ? Pourquoi ?



f) L'assureur s'intéresse à la queue à droite de la distribution et souhaite savoir les quantiles de niveau 99 % et 99,5 % pour les distributions des montants. Donner une estimation de ces quantiles sous les modèles log-normal et Pareto ajustés en d). [*Astuce : Observer qu'il est possible d'inverser explicitement la fonction de répartition de la Pareto, et que les quantiles de la log-normale peuvent être exprimés en termes de quantiles de la loi normale.*]

Réponses

4.1 a) $\bar{X}$ b) $\bar{X}/(1-\bar{X})$ c) $\bar{X}$ d) $\bar{X}$ e) $\bar{X}-1$

4.2 b) $\hat{\mu} = \ln\left\{\frac{\bar{Y}_n^2}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\right\}$ et $\hat{\sigma}^2 = \ln\left\{\frac{\frac{1}{n}\sum_{i=1}^n Y_i^2}{\bar{Y}_n^2}\right\}$

4.3 0,2381.

4.4 a) $\hat{\theta} = 2\bar{X} - 1$ b) 9,5

4.5 $\hat{\mu} = 3{,}241$, $\hat{\sigma} = 0{,}4$ et $\widehat{\Pr(X > 30)} = 0{,}3446$

4.6 $\hat{\tau} = 2$ et $\hat{\theta} = 200$

4.7 13,75

4.8 a) $\bar{X}$ b) $-n/\ln(X_1 \cdots X_n)$ c) $\bar{X}$ d) $\text{med}(X_1, \ldots, X_n)$ e) $X_{(1)}$

4.9 a) $E[X] = \mu + \lambda^{-1}$, $\text{var}[X] = \lambda^{-2}$

b) $\hat{\mu} = X_{(1)}$, $\hat{\lambda} = n/\sum_{i=1}^{n}(X_i - X_{(1)})$

4.11 $\hat{\mu} = \bar{X}$, $\hat{\lambda} = n/\sum_{i=1}^{n}(X_i^{-1} - \bar{X}^{-1})$

4.12 a) $\text{E}[X_1] = \theta/(\alpha - 1)$ et $\text{var}(X_1) = \alpha\theta^2/\{(\alpha - 1)^2(\alpha - 2)\}$
b) $\hat{\alpha} = 2S_n^2/(S_n^2 - \bar{X}_n^2)$ et $\hat{\theta} = \bar{X}_n(S_n^2 + \bar{X}_n^2)/(S_n^2 - \bar{X}_n^2)$

4.13 $\hat{a} = \min(X_1, \ldots, X_n)$ et $\hat{b} = \max(X_1, \ldots, X_n)$

4.14 a) $\{\max(X_1, \ldots, X_n) - 1\}/2$ b) $\{\max(X_1, \ldots, X_n)\}^2/12$

4.15 a) $\tilde{\theta} = \bar{X}$, $\hat{\theta} = \min(\bar{X}, 1/2)$

b) $\text{EQM}(\tilde{\theta}) = \theta(1-\theta)/n$, $\text{EQM}(\hat{\theta}) = \sum_{y=0}^{[n/2]}(y/n - \theta)^2\binom{n}{y}\theta^y(1-\theta)^{n-y} + \sum_{y=[n/2]+1}^{n}(1/2 - \theta)^2\binom{n}{y}\theta^y(1-\theta)^{n-y}$

c) $\text{EQM}(\hat{\theta}) \leq \text{EQM}(\tilde{\theta})$

4.16 b) $\hat{\mu} = 6.969$ et $\hat{\sigma}^2 = 1.112$ c) $\hat{\alpha} = 3.922$ et $\hat{\theta} = 5414.77$ d) Log-normale : 114 374,2, Pareto : 115 004,2 f) 11 296,76, 14 166,68, 12 720,54 et 16 537,1

5 Estimation par intervalle

5.1 Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n tiré d'une distribution Gamma avec paramètres $\alpha = 2$ et β inconnu.

a) Montrer que $T = 2(X_1 + \cdots + X_n)/\beta$ est un pivot et que sa distribution est khi-carrée avec $4n$ degrés de liberté. [Astuce : utiliser les fonctions génératrices des moments.]

b) Utiliser le pivot T pour construire un intervalle de confiance bilatéral de niveau 95 % pour β.

c) Si la moyenne échantillonnale est de $\bar{x} = 5{,}6$ et que $n = 5$, utiliser le résultat en b) pour donner un intervalle de confiance de niveau 95 % pour β.

5.2 La valeur observée de la moyenne empirique $\bar{X}$ d'un échantillon aléatoire de taille 20 tiré d'une $N(\mu, 80)$ est 81,2. Déterminer un estimateur par intervalle de niveau 95 % pour μ.

5.3 Soit $\bar{X}$ la moyenne d'un échantillon aléatoire de taille n d'une distribution normale de moyenne μ inconnue et de variance 9. Trouver la valeur n tel que, approximativement, $\Pr[\bar{X} - 1 < \mu < \bar{X} + 1] = 0{,}90$.

5.4 Un échantillon aléatoire comptant 17 observations d'une distribution normale de moyenne et de variance inconnues a donné $\bar{x} = 4{,}7$ et $s^2 = 5{,}76$. Trouver des intervalles de confiance à 90 % pour μ et pour σ^2.

5.5 Lors d'une très sérieuse et importante analyse statistique de la taille des étudiantes en sciences et génie à l'Université Laval, on a mesuré un échantillon aléatoire d'étudiantes en actuariat et un autre en génie civil. Les résultats obtenus se trouvent résumés dans le tableau ci-dessous. On suppose que les deux échantillons aléatoires sont indépendants et que la taille des étudiantes est distribuée selon une loi normale.

Quantité	Actuariat	Génie civil
Taille de l'échantillon	15	20
Taille moyenne (en cm)	152	154
Variance (en cm^2)	101	112

a) Déterminer un intervalle de confiance à 90 % pour la taille moyenne des étudiantes de chacun des deux programmes en supposant que l'écart type de la distribution normale est 9 cm.

b) Répéter la partie a) en utilisant plutôt les variances des échantillons.

c) On suppose que les variances des deux populations sont égales. Y a-t-il une différence significative, avec un niveau de confiance de 90 %, entre la taille des étudiantes en actuariat et celles en génie civil ?

d) Déterminer un intervalle de confiance à 90 % pour la variance de la taille des étudiantes en actuariat.

e) On suppose que les moyennes des deux populations sont égales. La différence observée entre les variances dans la taille des étudiantes des deux programmes est-elle significative ? Utiliser un niveau de confiance de 90 %.

5.6 Soit $X_1, \ldots, X_n$ un échantillon aléatoire tiré d'une population normale de moyenne et variance inconnues. Développer la formule d'un estimateur par intervalle de niveau $1 - \alpha$ pour σ, l'écart type de la distribution normale.

5.7 Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire d'une distribution normale de moyenne μ et de variance $\sigma^2 = 25$. Déterminer la taille de l'échantillon nécessaire pour que la longueur de l'intervalle de confiance de niveau 0,90 pour la moyenne ne dépasse pas 0,05.

5.8 Soit S_n^2 la variance échantillonnale d'un échantillon aléatoire de taille n issu d'une distribution $\mathcal{N}(\mu, \sigma^2)$ où μ et σ^2 sont des paramètres inconnus. On sait que $Y = (n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$. Soit $g(y)$ la fonction de densité de Y et $G(y)$ la fonction de répartition. Soit a et b des constantes telles que $((n-1)s^2/b, (n-1)s^2/a)$ est un intervalle de confiance de niveau $1 - \alpha$ pour σ^2. La longueur de cet intervalle est donc $(n-1)s^2(b-a)/(ab)$. Démontrer que la longueur de l'intervalle de confiance est minimale si a et b satisfont la condition $a^2 g(a) = b^2 g(b)$. (*Astuce* : minimiser la longueur de l'intervalle sous la contrainte que $G(b) - G(a) = 1 - \alpha$.)

5.9 Le *Scholastic Assessment Test* (SAT) est un test d'aptitudes standardisé largement utilisé dans les admissions aux universités américaines. Les résultats au test, qui avaient lentement diminués au fil des ans depuis son implantation, ont commencé à remonter. Initialement, un résultat de 500 était considéré dans la moyenne. En 2005, les résultats moyens étaient approximativement de 508 pour le test de langue et 520 pour le test de mathématiques. Un échantillon aléatoire de résultats de 20 étudiants d'une école secondaire a produit les moyennes et écart-types listés ci-dessous. On suppose que les résultats aux tests sont normalement distribués.

	Langue	Mathématiques
Moyenne échantillonnale	505	495
Écart-type échantillonnal	55	70

a) Trouver un intervalle de confiance bilatéral à 90 % pour la moyenne des résultats du test de langue des étudiants de l'école secondaire.

b) Est-ce que l'intervalle trouvé en a) inclut la valeur 508, la moyenne exacte du test de langue en 2005 ? Que peut-on en conclure ?

c) Trouver un intervalle de confiance bilatéral à 90 % pour la moyenne des résultats du test de mathématiques des étudiants de l'école secondaire. Est-ce que l'intervalle inclut la valeur 520, la moyenne du test de mathématiques en 2005 ? Que peut-on en conclure ?

d) Peut-on utiliser la méthode discutée en classe pour un construire un intervalle de confiance bilatéral pour la différence entre les moyennes des deux tests (langue et mathématiques), en supposant que leur variance est égale ? Expliquer.

e) Construire un intervalle de confiance bilatéral de niveau 90 % pour la variance σ^2 des résultats au test de mathématiques. Comment peut-on construire un intervalle de confiance bilatéral à 90 % pour σ ? Expliquer.

5.10 Le cuivre solide produit par frittage (chauffage sans fusion) d'une poudre dans des conditions environnementales spécifiques est ensuite mesuré pour la porosité (la fraction volumique due aux vides) en laboratoire. Un échantillon de $n = 4$ mesures de porosité indépendantes a une moyenne $\bar{x}_n = 0{,}22$ et une variance $s_n^2 = 0{,}001$. Un second laboratoire répète le même processus et obtient $m = 5$ mesures de porosité indépendantes avec $\bar{y}_m = 0{,}17$ and $s_m^2 = 0{,}002$.

a) Énumérer les hypothèses nécessaires pour construire un intervalle de confiance bilatéral *exact* pour la différence de moyennes entre les populations des deux laboratoires.

b) Construire un intervalle de confiance bilatéral *exact* à 95 % pour la différence des moyennes des populations des deux laboratoires.

c) Est-ce que la différence entre les moyennes des deux laboratoires est significative ?

5.11 Dans une étude sur la relation entre l'ordre de naissance et la réussite scolaire, un chercheur a trouvé que 126 étudiants gradués sur un échantillon de 180 étaient enfants ainés. Dans un échantillon de 100 personnes non graduées d'âge et de situation socio-économique comparable, le nombre d'ainés étaient de 54.

a) Construire un intervalle de confiance bilatéral approximatif à 90 % pour la différence entre les proportions d'ainés pour les deux populations desquelles proviennent les échantillons.

b) Basé sur l'intervalle en a), est-ce que la différence entre les proportions d'ainés des deux populations semble significative ?

c) Si les chercheurs souhaitent questionner un nombre égal n d'étudiants gradués et de personnes non graduées, quelle est la valeur minimale de n requise pour estimer la différence entre les proportions avec une erreur inférieure ou égale à ± 0.05 avec un niveau de confiance de 90 % ?

5.12 Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une distribution exponentielle avec densité

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

a) Écrire la vraisemblance, la log-vraisemblance et l'EMV de λ.

b) Calculer l'information de Fisher $I(\lambda)$.

c) Si $m = 100$ et $\bar{x} = 105{,}2$, utiliser la distribution limite de l'EMV pour obtenir un intervalle de confiance approximatif à 95% pour λ.

d) Basé sur l'intervalle trouvé en c), obtenir un intervalle de confiance approximatif de niveau 95 % pour la probabilité $\Pr[X > 300]$, où X est indépendant de $X_1, \ldots, X_n$ et a la même distribution.

Réponses

5.2 (77,28, 85,12)

5.3 24 ou 25

5.4 $\mu \in (3{,}7, 5{,}7)$ et $\sigma^2 \in (3{,}50, 11{,}58)$

5.5 a) $148{,}18 < \mu_1 < 155{,}82$ et $150{,}69 < \mu_2 < 157{,}31$

b) $147{,}43 < \mu_1 < 156{,}57$ et $149{,}91 < \mu_2 < 158{,}09$

c) $\mu_1 - \mu_2 \in -2 \pm 5{,}82$

d) $59{,}71 < \sigma_1^2 < 215{,}22$

e) $0{,}462 < \sigma_2^2/\sigma_1^2 < 2{,}502$

5.6 $(\sqrt{(n-1)S^2/b}, \sqrt{(n-1)S^2/a})$.

5.7 108 241

5.9 a) (483.734, 526.266)

b) Oui

c) (467.935, 522.065)

d) Non

e) (55.575, 95.929)

5.10 b) $(-0.01288, 0.11288)$

5.11 a) (0.06062, 0.25938)

b) Non

c) $n \geq 542$

5.12 c) [0.00764,0.01137] d) [0.03302,0.10099]

6 Tests

6.1 Soit une proportion p inconnue d'objets défectueux dans une grande population d'objets. On souhaite tester les hypothèses suivantes :

$$\mathcal{H}_0 : p = 0{,}2$$
$$\mathcal{H}_1 : p \neq 0{,}2.$$

On suppose qu'un échantillon aléatoire de $n = 20$ objets est tiré de la population. Soit X le nombre d'objets défectueux dans l'échantillon aléatoire, on considère un test d'hypothèse qui rejette l'hypothèse nulle si $X \geq 7$ ou $X \leq 1$.

a) Est-ce que les hypothèses sont simples ou composites ? Expliquer.

b) Quelle est la région critique du test ?

c) Calculer la taille du test et la probabilité de faire une erreur de type I.

d) Calculer la valeur de la fonction de puissance $\Pi(p)$ aux points

$$p \in \{0, 0{,}1, 0{,}2, 0{,}3, 0{,}4, 0{,}5, 0{,}6, 0{,}7, 0{,}8, 0{,}9, 1\}$$

et tracer la fonction de puissance en utilisant R ou Excel. Quel est le lien entre la fonction de puissance et les probabilités d'erreur de type I et II ?

e) Refaire (d) avec une taille d'échantillon de $n = 50$ et $n = 100$. Est-ce que le test semble bon ?

f) Suggérer une modification au test pour qu'il demeure significatif avec une taille d'échantillon différente de 20.

6.2 Soit X une variable aléatoire dont la fonction de densité de probabilité est

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 < x < 1.$$

On suppose que θ peut prendre exclusivement les valeurs $\theta = 1$ ou $\theta = 2$.

a) Trouver une statistique exhaustive pour le paramètre θ de cette distribution.

b) On teste l'hypothèse $\mathcal{H}_0 : \theta = 1$ versus $\mathcal{H}_1 : \theta = 2$ à partir d'un échantillon aléatoire X_1, X_2. Si la région critique est $C = \{(x_1, x_2); x_1 x_2 \geq 3/4\}$, calculer la probabilité de faire une erreur de type I (α) et la probabilité de faire une erreur de type II (β).

6.3 Soit $X_1, \ldots, X_n$ un échantillon aléatoire tiré d'une distribution Poisson avec paramètre λ inconnu. Soit λ_0 et λ_1 des valeurs spécifiques telles que $0 < \lambda_0 < \lambda_1$ et on souhaite tester les hypothèses suivantes :

$$\mathcal{H}_0 : \lambda = \lambda_0,$$
$$\mathcal{H}_1 : \lambda = \lambda_1.$$

[*On sait que si* $Z_1 \sim \mathcal{P}(\mu_1)$ *est indépendante de* $Z_2 \sim \mathcal{P}(\mu_2)$, *alors* $Z_1 + Z_2 \sim \mathcal{P}(\mu_1 + \mu_2)$.]

a) Montrer que le test optimal au seuil α rejette $\mathcal{H}_0$ quand $\bar{X}_n > c$. Spécifier la valeur de c.

b) Montrer que le test δ qui minimise $\alpha(\delta) + \beta(\delta)$ rejette $\mathcal{H}_0$ quand $\bar{X}_n > c^\star$. Trouver la valeur de $c^\star$.

c) On suppose que $n = 20$, $\lambda_0 = 1/20$ et $\lambda_1 = 1/10$. Calculer la valeur de la constante c en a) quand $\alpha = 0{,}08$ et calculer les probabilités d'erreur de type I et II.

d) On suppose que $n = 20$, $\lambda_0 = 1/20$ et $\lambda_1 = 1/10$. Calculer la valeur de $c^\star$ en b) et déterminer la valeur minimale que peut atteindre $\alpha(\delta) + \beta(\delta)$. Quelles sont les probabilités d'erreur de type I et II ?

6.4 On suppose que la durée de vie d'un pneu en kilomètres a une distribution normale de moyenne 30 000 et d'écart type 5 000. Le fabricant du nouveau pneu *Super Endurator X24* prétend que la durée de vie moyenne de ce pneu est bien supérieure à 30 000 km. Afin de vérifier les prétentions du fabricant, on testera $\mathcal{H}_0 : \mu \leq 30\,000$ versus la contre-hypothèse $\mathcal{H}_1 : \mu > 30\,000$ à partir de n observations indépendantes $x_1, \ldots, x_n$. On rejettera $\mathcal{H}_0$ si $\bar{x} \geq c$. Trouver les valeurs de n et de c de sorte que la probabilité de faire une erreur de type I en $\mu = 30\,000$ est 0,01 et que la probabilité de faire une erreur de type II en $\mu = 35\,000$ est 0,02.

6.5 On suppose que le poids en grammes des bébés à la naissance au Canada est distribué selon une loi normale de moyenne $\mu = 3\,315$ et de variance $\sigma^2 = 525^2$, garçons et filles confondus. Soit X le poids d'une fillette née au Québec. On suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$.

a) Donner l'expression de la statistique du test $\mathcal{H}_0 : \mu_X \leq 3\,315$ versus $\mathcal{H}_1 : \mu_X > 3\,315$ (les bébés sont en moyenne plus gros au Québec) si $n = 11$ et $\alpha = 0{,}01$. (La valeur de σ_X^2 est inconnue ici.)

b) Calculer la valeur de la statistique et tirer une conclusion si l'échantillon de poids de 11 fillettes nées au Québec est le suivant :

$$3\,119, \quad 2\,657, \quad 3\,459, \quad 3\,629, \quad 3\,345, \quad 3\,629,$$
$$3\,515, \quad 3\,856, \quad 3\,629, \quad 3\,345, \quad 3\,062.$$

c) Énoncer la statistique du test et la région critique du test $\mathcal{H}_0 : \sigma_X^2 \geq 525^2$ versus $\mathcal{H}_1 : \sigma_X^2 < 525^2$ (moins de variation dans le poids des bébés nés au Québec) si $\alpha = 0{,}05$.

d) Calculer la statistique du test à partir des données de la partie b). Quelle est la conclusion à tirer de ce résultat ?

6.6 Soit Y le poids en grammes d'un garçon né au Québec et supposons que $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. On a les observations suivantes :

$$\begin{array}{llllll} 4\,082, & 3\,686, & 4\,111, & 3\,686, & 3\,175, & 4\,139, \\ 3\,686, & 3\,430, & 3\,289, & 3\,657, & 4\,082. & \end{array}$$

Refaire les questions de l'exercice 6.5. Les réponses obtenues dans ces deux exercices suggèrent-elles d'autres hypothèses à explorer ? Faire les tests appropriés le cas échéant.

6.7 Parmi les statistiques relevées par l'Organisation mondiale de la santé (OMS) on compte la concentration en $\mu g/m^3$ de particules en suspension dans l'air. Soit X et Y les concentrations en $\mu g/m^3$ de particules en suspension dans l'air aux centres-villes de Melbourne (Australie) et Houston (Texas), respectivement. On suppose que X et Y sont normalement distribuées. À partir de $n = 13$ observations de la variable aléatoire X et $m = 16$ observations de la variable aléatoire Y, on souhaite tester $\mathcal{H}_0 : \mu_X \geq \mu_Y$ versus $\mathcal{H}_1 : \mu_X < \mu_Y$.

a) Définir la statistique du test et la région critique en supposant égales les variances des distributions de X et Y. Utiliser un seuil de signification de 5 %.

b) Si $\bar{x} = 72{,}9$, $s_X = 25{,}6$, $\bar{y} = 81{,}7$ et $s_Y = 28{,}3$, quelle est la conclusion pour ce test ?

c) Calculer la valeur p de ce test. Est-elle conforme à la conclusion en b) ?

d) Tester si l'hypothèse de variances égales faite en a) est valide avec un niveau de confiance de 90 %.

6.8 Un fabricant de dentifrice prétend que 75 % de tous les dentistes recommandent son produit à ses patients. Sceptique, un groupe de protection des consommateurs décide de tester $\mathcal{H}_0 : \theta = 0{,}75$ contre $\mathcal{H}_1 : \theta \neq 0{,}75$, où θ est la proportion de dentistes recommandant le dentifrice en question. Un sondage auprès de 390 dentistes a révélé que 273 d'entre eux recommandent effectivement ce dentifrice.

a) Quelle est la conclusion du test avec un seuil de signification de $\alpha = 0{,}05$?

b) Quelle est la conclusion du test avec un seuil de signification de $\alpha = 0{,}01$?

c) Quel est le seuil observé du test ?

6.9 Soit θ la proportion de bonbons rouges dans une boîte de Smarties. On prétend que $\theta = 0{,}20$.

a) Définir la statistique de test et la région critique avec un seuil de signification de 5 % pour le test

$$\begin{aligned} \mathcal{H}_0 &: \theta = 0{,}2 \\ \mathcal{H}_1 &: \theta \neq 0{,}2. \end{aligned}$$

b) Pour faire le test, les 20 membres de la section locale des Amateurs de Smarties Associés (ASA) ont chacun compté le nombres de bonbons rouges dans une boîte

de 50 grammes de Smarties. Ils ont obtenu les proportions suivantes :

$$\frac{8}{56}, \frac{13}{55}, \frac{12}{58}, \frac{13}{56}, \frac{14}{57}, \frac{5}{54}, \frac{14}{56}, \frac{15}{57}, \frac{11}{54}, \frac{13}{55},$$

$$\frac{10}{57}, \frac{8}{59}, \frac{10}{54}, \frac{11}{55}, \frac{12}{56}, \frac{11}{57}, \frac{6}{54}, \frac{7}{58}, \frac{12}{58}, \frac{14}{58}.$$

Si chaque membre de l'ASA fait le test mentionné en a), quelle proportion des membres rejette l'hypothèse $\mathcal{H}_0$?

c) En supposant vraie l'hypothèse $\mathcal{H}_0$, à quelle proportion de rejets de l'hypothèse $\mathcal{H}_0$ peut-on s'attendre ?

d) Pour chacun des ratios en b) on peut construire un intervalle de confiance à 95 % pour θ. Quelle proportion de ces intervalles de confiance contiennent $\theta = 0{,}20$?

e) Si les 20 résultats en b) sont agrégés de sorte que l'on a compté un total de 219 bonbons rouges parmi 1 124 Smarties, rejette-t-on l'hypothèse $\mathcal{H}_0$, toujours avec $\alpha = 0{,}05$?

6.10 Lors d'un sondage mené auprès de 800 adultes dont 605 non-fumeurs, on a posé la question suivante : *Devrait-on introduire une nouvelle taxe sur le tabac pour aider à financer le système de santé au pays ?* Soit θ_1 et θ_2 la proportion de non-fumeurs et de fumeurs, respectivement, qui ont répondu par l'affirmative à cette question. Les résultats du sondage sont les suivants : $x_1 = 351$ non-fumeurs ont répondu oui, contre $x_2 = 41$ fumeurs.

a) Tester l'hypothèse nulle que $\theta_1 = \theta_2$ versus la contre-hypothèse $\theta_1 \neq \theta_2$ avec un seuil de signification de 5 %.

b) Trouver un intervalle de confiance à 95 % pour $\theta_1 - \theta_2$. Cet intervalle permet-il d'obtenir la même conclusion qu'en a) ?

c) Trouver un intervalle de confiance à 95 % pour la proportion de la population totale en faveur de l'introduction d'une nouvelle taxe sur le tabac.

6.11 Télécharger le fichier `contents.csv` sur le site de cours. Il contient les pertes de biens, dues à un incendie, de plus d'un million de couronnes danoises provenant de réclamations faites à la compagnie de réassurance Copenhagen Re entre 1980 et 1990. Les données peuvent être importées en R comme suit, après avoir changé l'environnement de travail de R au dossier dans lequel vous avez enregistré le fichier `contents.csv` :

```
ct <- read.csv("donnees/contents.csv",header=TRUE,row.names=1)
attach(ct)
data <- Contents
```

Les données peuvent aussi être importées par RStudio en cliquant sur "Import Dataset". On se rappelle également que la distribution exponentielle est un cas spécial de la distribution Gamma quand le paramètre α égal 1.

a) Ajuster la distribution Gamma aux données et construire un intervalle de confiance à 95% pour α. Est-ce qu'il contient la valeur $\alpha = 1$? Que peut-on en conclure ?

b) Tester l'hypothèse que $\alpha = 1$ à un niveau de 5% en utilisant le test de Wald. Comparer la conclusion avec celle en (a).

c) Tester l'hypothèse que le modèle exponentiel est une simplification adéquate du modèle Gamma en utilisant un test du rapport de vraisemblance au niveau 5%. Comparer la conclusion avec celle en (b).

6.12 Le tableau de contingence suivant montre le destin des passagers du Titanic en fonction de la classe de tarification du billet, représentant le statut socio-économique.

```
##          Class
## Survived 1st 2nd 3rd Crew
##      No  122 167 528  673
##      Yes 203 118 178  212
```

a) Peut-on conclure que la probabilité de survie au naufrage du Titanic était différente selon la classe tarifaire au seuil de 5 % ?

b) Avec R, calculer le seuil observé du test utilisé en a).

c) Avec R, visualiser les probabilités de survie estimées avec un diagramme en mosaïque.

Réponses

6.2 a) $\prod_{i=1}^{n} X_i$

b) $\alpha = 0{,}034$, $\beta = 0{,}886$

6.4 $n = 19$ ou 20, $c = 32\,658$

6.5 b) $t = 0{,}699 < t_{10,0,01}$ d) $y = 4{,}104 > \chi^2_{10,0,05}$

6.6 b) $t = 4{,}028 > t_{10,0,01}$ d) $y = 4{,}223 > \chi^2_{10,0,95}$ e) tests sur les moyennes et les variances

6.7 a) $T = [(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)] / \sqrt{((n-1)S_X^2 + (m-1)S_Y^2)(n^{-1} + m^{-1})/(n+m-2)}$, $T \sim t(n+m-2)$ b) $t = -0{,}838 > t_{0,05}(27)$ c) $p = 0{,}2047$ d) $f = 0{,}8311 < f_{0,05}(12,15)$

6.8 a) $z = -2{,}28$ c) 0,0226

6.9 a) $Z = (\hat{\theta} - 0{,}20)/\sqrt{0{,}20(1-0{,}20)/n}$ b) 0 % c) 0,05 d) 100 % e) $p = 0{,}6654$

6.10 a) $z = 10{,}45 > 1{,}96$ b) (0,3005, 0,4393) c) (0,4555, 0,5246)

6.12 a) oui b) $5.0276183 \times 10^{-41}$

A Tables

A.1 Table de la loi normale

Le tableau donne les valeurs de la fonction de répartition Φ de la distribution $\mathcal{N}(0,1)$, évaluée à z.

	z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

A.2 Table de quantiles de la loi khi carré

Le tableau donne $\chi^2_\alpha(\nu)$, le quantile supérieur de niveau α de la loi khi carré avec ν degrés de liberté, α est donné dans les colonnes, et ν est donné dans les lignes. Précision : Si $X \sim \chi^2(\nu)$, alors $\Pr\{X > \chi^2_\alpha(\nu)\} = \alpha$.

	Queue de gauche					Queue de droite					
ν	0.99500	0.99000	0.97500	0.95000	0.90000	0.10000	0.05000	0.02500	0.01000	0.00500
1	0.00004	0.00016	0.00098	0.00393	0.01579	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	40.25602	43.77297	46.97924	50.89218	53.67196
40	20.70654	22.16426	24.43304	26.50930	29.05052	51.80506	55.75848	59.34171	63.69074	66.76596
50	27.99075	29.70668	32.35736	34.76425	37.68865	63.16712	67.50481	71.42020	76.15389	79.48998
60	35.53449	37.48485	40.48175	43.18796	46.45889	74.39701	79.08194	83.29767	88.37942	91.95170
70	43.27518	45.44172	48.75756	51.73928	55.32894	85.52704	90.53123	95.02318	100.42518	104.21490
80	51.17193	53.54008	57.15317	60.39148	64.27784	96.57820	101.87947	106.62857	112.32879	116.32106
90	59.19630	61.75408	65.64662	69.12603	73.29109	107.56501	113.14527	118.13589	124.11632	128.29894
100	67.32756	70.06489	74.22193	77.92947	82.35814	118.49800	124.34211	129.56120	135.80672	140.16949

A.3 Table de quantiles de la loi t

Le tableau donne $t_{\nu,\alpha}$, le quantile supérieur de niveau α de la loi de Student avec ν degrés de liberté, α est donné dans les colonnes, ν est donné dans les lignes. Précision : Si $T \sim t_{(\nu)}$, alors $\Pr\{T > t_{\nu,\alpha}\} = \alpha$.

	α				
ν	0.100	0.050	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

A.4 Table de quantiles de la loi Fisher–Snedecor

Le tableau donne le quantile supérieur de niveau $\alpha = 0.05$ de la distribution $\mathsf{F}(\nu_1,\nu_2)$; ν_1 est donné dans les colonnes, ν_2 est donné dans les lignes.

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	9	10	11	12
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	242.98	243.91
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.28	2.25
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.24	2.20
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.17	2.13
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.14	2.10
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.11	2.08
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	2.10	2.07

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	9	10	11	12
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	2.09	2.06
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	2.08	2.05
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	2.07	2.03
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10	2.06	2.02
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	2.05	2.02
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08	2.04	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
41	4.08	3.23	2.83	2.60	2.44	2.33	2.24	2.17	2.12	2.07	2.03	2.00
42	4.07	3.22	2.83	2.59	2.44	2.32	2.24	2.17	2.11	2.06	2.03	1.99
43	4.07	3.21	2.82	2.59	2.43	2.32	2.23	2.16	2.11	2.06	2.02	1.99
44	4.06	3.21	2.82	2.58	2.43	2.31	2.23	2.16	2.10	2.05	2.01	1.98
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	2.01	1.97
46	4.05	3.20	2.81	2.57	2.42	2.30	2.22	2.15	2.09	2.04	2.00	1.97
47	4.05	3.20	2.80	2.57	2.41	2.30	2.21	2.14	2.09	2.04	2.00	1.96
48	4.04	3.19	2.80	2.57	2.41	2.29	2.21	2.14	2.08	2.03	1.99	1.96
49	4.04	3.19	2.79	2.56	2.40	2.29	2.20	2.13	2.08	2.03	1.99	1.96
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95
51	4.03	3.18	2.79	2.55	2.40	2.28	2.20	2.13	2.07	2.02	1.98	1.95
52	4.03	3.18	2.78	2.55	2.39	2.28	2.19	2.12	2.07	2.02	1.98	1.94
53	4.02	3.17	2.78	2.55	2.39	2.28	2.19	2.12	2.06	2.01	1.97	1.94
54	4.02	3.17	2.78	2.54	2.39	2.27	2.18	2.12	2.06	2.01	1.97	1.94
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01	1.97	1.93
56	4.01	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.05	2.00	1.96	1.93
57	4.01	3.16	2.77	2.53	2.38	2.26	2.18	2.11	2.05	2.00	1.96	1.93
58	4.01	3.16	2.76	2.53	2.37	2.26	2.17	2.10	2.05	2.00	1.96	1.92
59	4.00	3.15	2.76	2.53	2.37	2.26	2.17	2.10	2.04	2.00	1.96	1.92
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
61	4.00	3.15	2.76	2.52	2.37	2.25	2.16	2.09	2.04	1.99	1.95	1.91
62	4.00	3.15	2.75	2.52	2.36	2.25	2.16	2.09	2.03	1.99	1.95	1.91
63	3.99	3.14	2.75	2.52	2.36	2.25	2.16	2.09	2.03	1.98	1.94	1.91
64	3.99	3.14	2.75	2.52	2.36	2.24	2.16	2.09	2.03	1.98	1.94	1.91

B Solutions

Chapitre 1

1.1 a) Les progrès en médecine ont affecté grandement la probabilité de détection précoce du cancer du sein, donc la probabilité de survie. Il y a donc une tendance à la baisse, et ces données ne sont pas identiquement distribuées dans le temps.

b) Puisque chaque lancer est indépendant et identiquement distribué, les 20 lancers sont considérés comme un échantillon aléatoire de taille 20 de lancers de dé.

1.2 a) Il faudrait utiliser une distribution continue non négative sur l'intervalle [0,∞) étant donné que la perte financière est nécessairement positive et qu'il n'y a pas de borne supérieure naturelle. Des bons exemples de lois pourraient être : gamma, lognormale, Pareto.

b) La distribution adéquate est la loi binomiale négative. Considérant que chaque sac est indépendant des autres et a une probabilité p d'être contaminé, la distribution est une loi binomiale négative avec paramètres $r = 15$ et p inconnu.

c) La loi binomiale est appropriée. Considérant que chaque sac est indépendant des autres et a une probabilité p d'être contaminé, le nombre total de sacs contaminés suit une loi binomiale avec paramètres de taille $n = 15$ et probabilité p inconnu.

d) Le contexte est conforme à une distribution normale. La majorité des conducteurs vont tendre vers une vitesse moyenne malgré que certains vont conduire beaucoup plus vite ou lentement ; une distribution en forme de cloche semble donc appropriée. La moyenne μ et la variance σ^2 de la vitesse réelle des véhicules à cet endroit précis sur l'autoroute sont inconnues.

e) Le contexte suggère une loi Normale. Les câbles devraient avoir une résistance près de μ, soit celle donnée par le producteur. L'écart-type σ dépendra de la qualité de la précision de la fabrication, et est inconnu.

1.3 a) On a directement $f(x) = F'(x) = \alpha\beta x^{\alpha-1}e^{-\beta x^\alpha}$.

b) On a

$$\begin{aligned} E[X] &= \int_0^\infty x f(x)\,dx \\ &= \int_0^\infty \alpha\beta x^\alpha e^{-\beta x^\alpha}\,dx. \end{aligned}$$

On effectue le changement de variable $y = \beta x^\alpha$, d'où $dy = \boldsymbol{\alpha\beta x^{\alpha-1}}\, \boldsymbol{dx}$ et $x = \left(\frac{y}{\beta}\right)^{\frac{1}{\alpha}}$ et donc

$$\begin{aligned} E[X] &= \int_0^\infty \boldsymbol{\alpha\beta x^{\alpha-1}} x e^{-\beta x^\alpha}\, \boldsymbol{dx} \\ &= \frac{1}{\beta^{1/\alpha}} \int_0^\infty y^{1/\alpha} e^{-y}\, dy \\ &= \frac{\Gamma(1+\frac{1}{\alpha})}{\beta^{1/\alpha}} \int_0^\infty \frac{1}{\Gamma(1+\frac{1}{\alpha})} y^{1/\alpha} e^{-y}\, dy \\ &= \frac{\Gamma(1+\frac{1}{\alpha})}{\beta^{1/\alpha}} \end{aligned}$$

puisque l'intégrande ci-dessus est la fonction de densité de probabilité d'une loi gamma de paramètre de forme $\alpha = 1 + \frac{1}{\alpha}$ et de paramètre d'échelle $\beta = 1$. En procédant exactement de la même façon, on trouve

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2 f(x)\, dx \\ &= \int_0^\infty \alpha\beta x^{\alpha+1} e^{-\beta x^\alpha}\, dx \\ &= \frac{1}{\beta^{2/\alpha}} \int_0^\infty y^{2/\alpha} e^{-y}\, dy \\ &= \frac{\Gamma(1+\frac{2}{\alpha})}{\beta^{2/\alpha}}. \end{aligned}$$

Par conséquent,

$$\text{var}[X] = \frac{\Gamma(1+\frac{2}{\alpha}) - \Gamma(1+\frac{1}{\alpha})^2}{\beta^{2/\alpha}}.$$

1.4 Soit X une variable aléatoire avec distribution Beta(α,β). Alors,

$$\begin{aligned} \mathrm{E}(X) &= \int_0^1 x \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathrm{d}x \\ &= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha}(1-x)^{\beta-1} \mathrm{d}x. \end{aligned}$$

Cette intégrale peut être résolue en reconnaissant la forme similaire à une distribution Beta($\alpha+1$,β) intégrée sur son domaine, mais avec les mauvaises constantes. On divise et multiplie par les constantes nécessaires pour trouver une intégrale avec

valeur de 1.

$$\begin{aligned}
\mathrm{E}(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha+1+\beta)}\int_0^1 \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta-1}\mathrm{d}x \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+1+\beta)}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \frac{\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \frac{\alpha}{\alpha+\beta}.
\end{aligned}$$

De plus,

$$\begin{aligned}
\mathrm{E}\{X^2\} &= \int_0^1 x^2\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x \\
&= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha+1}(1-x)^{\beta-1}\mathrm{d}x \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)}\frac{\Gamma(\alpha+2)}{\Gamma(\alpha+2+\beta)}\int_0^1 \frac{\Gamma(\alpha+2+\beta)}{\Gamma(\alpha+2)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta-1}\mathrm{d}x \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+2+\beta)}\frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} \\
&= \frac{\Gamma(\alpha+\beta)}{(\alpha+\beta+1)\Gamma(\alpha+\beta+1)}\frac{(\alpha+1)\Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \frac{\Gamma(\alpha+\beta)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)}\frac{(\alpha+1)\alpha\Gamma(\alpha)}{\Gamma(\alpha)} \\
&= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}.
\end{aligned}$$

La variance peut alors être calculée ainsi :

$$\begin{aligned}
\mathrm{var}(X) &= \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2 \\
&= \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha}{\alpha+\beta}\left[\frac{\alpha+1}{\alpha+\beta+1} - \frac{\alpha}{\alpha+\beta}\right] \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}$$

1.5 Soit X une variable aléatoire Poisson avec paramètre $\lambda > 0$. Sa fonction génératrice des moments, pour tout $t \in \mathbb{R}$, est donnée par

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\{\lambda e^t\}^k}{k!}.$$

L'équation de droite est la série de Taylor de $\exp(\lambda e^t)$. Ainsi,

$$M_X(t) = e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(e^t - 1)\}.$$

La ke dérivée de M évaluée à $t = 0$ donne le ke moment de X. Ainsi,

$$\mathrm{E}(X) = M'_X(0) = \exp\{\lambda(e^t - 1)\} \times \lambda e^t\big|_{t=0} = \lambda$$

de façon similaire,

$$\mathrm{E}(X^2) = M''_X(0) = \exp\{\lambda(e^t - 1)\} \times \lambda^2 e^{2t} + \exp\{\lambda(e^t - 1)\} \times \lambda e^t\Big|_{t=0} = \lambda^2 + \lambda$$

Alors,

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \{\mathrm{E}(X)\}^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

1.6 a) Le support de Y est [0,2], donc $F(y) = 0$ pour $y < 0$ et $F(y) = 1$ pour $y > 2$. Pour $0 \leq y < 1$,

$$F(y) = \Pr[Y \leq y] = \int_0^y x\mathrm{d}x = \frac{y^2}{2}.$$

Pour $1 \leq y < 2$,

$$F(y) = \Pr[Y \leq y] = \int_0^1 x\mathrm{d}x + \int_1^y (2 - x)\mathrm{d}x = 1 - \frac{(2-y)^2}{2}.$$

Ainsi la fonction de répartition de Y est

$$F(y) = \begin{cases} 0, & y < 0 \\ y^2/2, & 0 \leq y < 1 \\ 1 - (2-y)^2/2, & 1 \leq y < 2 \\ 1, & y > 2. \end{cases}$$

b) On souhaite trouver la probabilité que $10000Y$ se situe entre 8 500 et 11 500 :

$$\begin{aligned} \Pr[8500 < 10000Y \leq 11500] &= \Pr[0.85 < Y \leq 1.15] \\ &= F(1.15) - F(0.85) \\ &= 1 - \frac{(2-1.15)^2}{2} - \frac{0.85^2}{2} = 0.2775. \end{aligned}$$

c) On désire trouver l'espérance du revenu, qui est de 2,10 \$ par litre, donc $\mathrm{E}[2.10 \times 10000Y] = 21000\,\mathrm{E}[Y]$. On a

$$\begin{aligned} \mathrm{E}[Y] &= \int_0^1 y^2\mathrm{d}y + \int_1^2 y(2-y)\mathrm{d}y \\ &= \frac{y^3}{3}\bigg|_0^1 + y^2 - \frac{y^3}{3}\bigg|_1^2 \\ &= \frac{1}{3} + 4 - \frac{2^3}{3} - 1 + \frac{1}{3} = 1. \end{aligned}$$

Ainsi, l'espérance de revenu mensuel est de 21 000 \$.

1.7 On doit trouver le point où la fonction $f(c) = E[(X-c)^2]$ atteint son minimum. Pour ce faire, il faut dériver par rapport à c. Or, $\frac{\mathrm{d}}{\mathrm{d}c} f(c) = -2E[X-c] = 0$ lorsque $E[X] - c = 0$, soit $c = E[X] = \mu$. De plus, il s'agit bien d'un minimum puisque $\frac{\mathrm{d}^2}{\mathrm{d}^2 c} f(c) = 2 > 0$ pour tout c.

1.8 a) On a

$$\begin{aligned} M_Y(t) &= E[e^{Yt}] \\ &= E[e^{(aX+b)t}] \\ &= E[e^{aXt}e^{bt}] \\ &= e^{bt}E[e^{aXt}] \\ &= e^{bt}M_X(at). \end{aligned}$$

b) On a

$$\begin{aligned} M_Y(t) &= E[e^{Yt}] \\ &= E[e^{(X_1+\cdots+X_n)t}] \\ &= E[e^{X_1 t}\cdots e^{X_n t}] \end{aligned}$$

et, par indépendance entre les variables aléatoires,

$$\begin{aligned} M_Y(t) &= E[e^{X_1 t}]\cdots E[e^{X_n t}] \\ &= \prod_{i=j}^{n} M_{X_j}(t). \end{aligned}$$

Si, en plus, les variables aléatoires $X_1, \ldots, X_n$ sont identiquement distribuées comme X, alors $M_Y(t) = (M_X(t))^n$.

1.9 La Loi faible des grands nombres indique que, lorsque $n \to \infty$, $\bar{X}_n \xrightarrow{P} \mathrm{E}[X]$, où $X_1, \ldots, X_n$ est une suite de variables aléatoires indépendantes et identiquement ditribuées. Cependant, le résultat tient seulement si $\mathrm{E}[X^2] < \infty$, ce qui n'est pas le cas ici puisque

$$\mathrm{E}[X^2] = \int_{10}^{\infty} \frac{200}{x^3} x^2 \mathrm{d}x = \int_{10}^{\infty} \frac{200}{x} \mathrm{d}x = 200 \ln(x)\big|_{10}^{\infty} = \infty.$$

Ainsi, il n'est pas possible d'appliquer la WLLN dans ce cas.

1.10 Si X a la même distribution que $X_1, \ldots, X_n$ et $\mathrm{var}(X) < \infty$, la loi faible des grands nombres implique que, quand $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge en probabilité vers $E[X]$. Puisque X suit une distribution Beta avec paramètres $\alpha = 1$ et $\beta = 4$, $E[X] = \alpha/(\alpha+\beta) = 1/5$ et

$$\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{4}{5^2 \times 6} = \frac{2}{75} < \infty.$$

Donc, la WLLN s'applique et $\bar{X}_n$ converge en probabilité vers $1/5$.

1.11 Puisque $X_{(1)}$ est la plus petite valeur de l'échantillon, on a que

$$\begin{aligned} F_{X_{(1)}}(x) &= \Pr[X_{(1)} \leq x] \\ &= 1 - \Pr[X_{(1)} > x] \\ &= 1 - \Pr[X_1 > x, X_2 > x, \ldots, X_n > x]. \end{aligned}$$

Or, les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et identiquement distribuées, d'où

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - \Pr[X_1 > x]\Pr[X_2 > x] \cdots \Pr[X_n > x] \\ &= 1 - (\Pr[X > x])^n \\ &= 1 - (1 - F_X(x))^n. \end{aligned}$$

1.12 On cherche la probabilité que la plus grande valeur de l'échantillon soit supérieure à 3, soit le complément de la probabilité que toutes les valeurs de l'échantillon soient inférieures à 3 :

$$\begin{aligned} \Pr[X_{(4)} > 3] &= 1 - \Pr[X_{(4)} \leq 3] \\ &= 1 - \Pr[X_1 \leq 3]\Pr[X_2 \leq 3]\Pr[X_3 \leq 3]\Pr[X_4 \leq 3] \\ &= 1 - (F_X(3))^4. \end{aligned}$$

Or, on aura reconnu en $f(x)$ la densité d'une loi exponentielle de paramètre $\lambda = 1$. Par conséquent, $F_X(x) = 1 - e^{-x}$ et $\Pr[X_{(4)} > 3] = 1 - (1 - e^{-3})^4$.

1.13 Soit m la médiane de la distribution. On cherche $\Pr[X_{(1)} > m]$. Avec le résultat de l'exercice 1.11,

$$\begin{aligned} \Pr[X_{(1)} > m] &= 1 - \Pr[X_{(1)} \leq m] \\ &= 1 - F_{X_{(1)}}(m) \\ &= 1 - (1 - (1 - F_X(m))^3) \\ &= (1 - F_X(m))^3 \\ &= \frac{1}{8}, \end{aligned}$$

car $F_X(m) = 1 - F_X(m) = 1/2$ par définition de la médiane. Le type de distribution ne joue donc aucun rôle dans cet exercice.

1.14 On a que X est distribuée uniformément sur $\{1, \ldots, 6\}$, d'où $F_X(x) = x/6$, $x = 1, \ldots, 6$. De l'exercice 1.11, on a que

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - (1 - F_X(x))^5 \\ &= 1 - \left(1 - \frac{x}{6}\right)^5 \end{aligned}$$

comme la taille d'échantillon est $n = 5$. Par conséquent, la fonction de masse de probabilité du minimum est

$$\begin{aligned}
\Pr[X_{(1)} = x] &= \lim_{y \to x^+} F_{X_{(1)}}(y) - \lim_{y \to x^-} F_{X_{(1)}}(y) \\
&= \Pr[X_{(1)} \leq x] - \Pr[X_{(1)} < x] \\
&= \Pr[X_{(1)} \leq x] - \Pr[X_{(1)} \leq (x-1)] \\
&= F_{X_{(1)}}(x) - F_{X_{(1)}}(x-1) \\
&= \left(1 - \frac{x-1}{6}\right)^5 - \left(1 - \frac{x}{6}\right)^5 \\
&= \left(\frac{7-x}{6}\right)^5 - \left(\frac{6-x}{6}\right)^5.
\end{aligned}$$

1.15 De l'exercice 1.11, on a

$$\begin{aligned}
F_{X_{(1)}}(x) &= 1 - (1 - F_X(x))^n \\
&= 1 - (e^{-\beta x^\alpha})^n \\
&= 1 - e^{-n\beta x^\alpha},
\end{aligned}$$

d'où $X_{(1)} \sim \text{Weibull}(\alpha, n\beta)$. Ainsi, la fonction de densité de probabilité du minimum de l'échantillon est

$$f_{X_{(1)}}(x) = n\beta\alpha x^{\alpha-1} e^{-n\beta x^\alpha}$$

et l'espérance est

$$E[X_{(1)}] = \frac{\Gamma(1 + 1/\alpha)}{(n\beta)^{\frac{1}{\alpha}}}.$$

1.16 À partir de la fonction de répartition, on a

$$\begin{aligned}
F_{X_{(1)},X_{(n)}}(x,y) &= \Pr[X_{(1)} \leq x, X_{(n)} \leq y] \\
&= \Pr[X_{(n)} \leq y] - \Pr[X_{(1)} > x, X_{(n)} \leq y] \\
&= \Pr[X_1 \leq y, \ldots, X_n \leq y] - \Pr[x < X_1 \leq y, \ldots, x < X_n \leq y] \\
&= (F(y))^n - (F(y) - F(x))^n, \quad \text{car iid.}
\end{aligned}$$

Donc, pour $x < y$,

$$f_{X_{(1)},X_{(n)}}(x,y) = \frac{\mathrm{d}F_{X_{(1)},X_{(n)}}(x,y)}{\mathrm{d}x\mathrm{d}y} = n(n-1)(F(y) - F(x))^{n-2} f(x) f(y).$$

1.17 On pose

$$r = y - x \text{ et } t = \frac{x+y}{2}$$

ce qui est équivalent à

$$x = -\frac{r}{2} + t \text{ et } y = \frac{r}{2} + t.$$

Le jacobien de la transformation de (x,y) vers (r,t) est

$$J = \begin{vmatrix} -1/2 & 1 \\ 1/2 & 1 \end{vmatrix} = -1.$$

On a donc que la densité conjointe de l'étendue R et de la mi-étendue T est

$$f_{R,T}(r,t) = n(n-1)\left(F_X\left(t+\frac{r}{2}\right) - F_X\left(t-\frac{r}{2}\right)\right)^{n-2} f_X\left(t+\frac{r}{2}\right) f_X\left(t-\frac{r}{2}\right),$$

pour $r > 0$ et $-\infty < t < \infty$.

1.18 Ici, $f_X(x) = 1$ pour $x \in (0,1)$ et $F_X(x) = x$, pour $x \in (0,1)$. Selon l'exercice 1.17, on trouve que

$$\begin{aligned} f_{R,T}(r,t) &= n(n-1)\{F_X(t+r/2) - F_X(t-r/2)\}^{n-2} f_X(t+r/2) f_X(t-r/2) \\ &= n(n-1)\{(t+r/2) - (t-r/2)\}^{n-2} \\ &= n(n-1)r^{n-2}, \end{aligned}$$

pour $0 < r < 1$ et $t + r/2 \in (0,1)$ et $t - r/2 \in (0,1)$.Ainsi,

$$\begin{aligned} t + \frac{r}{2} < 1 &\implies t < 1 - \frac{r}{2} \\ t - \frac{r}{2} > 0 &\implies t > \frac{r}{2} \\ &\implies \frac{r}{2} < t < 1 - \frac{r}{2}. \end{aligned}$$

Par conséquent,

$$\begin{aligned} f_R(r) &= n(n-1)r^{n-2} \int_{r/2}^{1-r/2} dt \\ &= n(n-1)r^{n-2}(1-r), \quad 0 < r < 1. \end{aligned}$$

1.19 Soit R l'étendue de l'échantillon aléatoire. Avec l'exercice 1.18, on sait que

$$f_R(x) = n(n-1)x^{n-2}(1-x), \text{ pour } x \in (0,1).$$

Par conséquent

$$\begin{aligned} \Pr\left[R \le \frac{1}{2}\right] &= \int_0^{1/2} f_X(x)\,dx \\ &= (4)(3)\int_0^{1/2} x^2(1-x)\,dx \\ &= \frac{5}{16}. \end{aligned}$$

1.20 On a que $X \sim$ Bêta(1,2), c'est-à-dire que $f_X(x) = 2(1-x)$, $0 < x < 1$. Soit X_1, X_2 un échantillon aléatoire tiré de cette densité. Par indépendance, on a

$$\begin{aligned} f_{X_1X_2}(x_1,x_2) &= f_{X_1}(x) f_{X_2}(x) \\ &= 4(1-x_1)(1-x_2). \end{aligned}$$

On cherche $\Pr[X_2 \geq 2X_1 \cup X_1 \geq 2X_2]$. Par définition,

$$\Pr[X_1 \geq 2X_2 \cup X_2 \geq 2X_1] = \iint_{\mathcal{R}} f_{X_1X_2}(x_1,x_2)\, dx_2\, dx_1,$$

où $\mathcal{R}$ est la région du domaine de définition de $f_{X_1X_2}$ telle que $x_1 > 2x_2$ ou $x_2 > 2x_1$. Cette région est représentée à la figure B.1. On a donc

$$\begin{aligned} \Pr[X_1 \geq 2X_2 \cup X_2 \geq 2X_1] &= 4\int_0^{1/2}\int_{2x_1}^1 (1-x_1)(1-x_2)\, dx_2\, dx_1 \\ &\quad + 4\int_0^{1/2}\int_{2x_2}^1 (1-x_1)(1-x_2)\, dx_1\, dx_2 \\ &= 4\int_0^{1/2}(1-x_1)\left(\frac{1}{2} - 2x_1 + 2x_1^2\right) dx_1 \\ &\quad + 4\int_0^{1/2}(1-x_2)\left(\frac{1}{2} - 2x_2 + 2x_2^2\right) dx_2 \\ &= \frac{7}{12}. \end{aligned}$$

1.21 Soit $T = (X_{(1)} + X_{(n)})/2$ la mi-étendue et R l'étendue. On sait que, pour $r > 0$ et $-\infty < t < \infty$,

$$f_{RT}(r,t) = n(n-1)r^{n-2}.$$

On doit calculer la densité marginale de T. Il faut voir que le domaine de R (et donc le domaine d'intégration) dépend indirectement de T.

Le domaine de T est d'abord $0 \leq t \leq 1$ comme $X_{(1)}$ et $X_{(n)}$ sont chacun au maximum 1. Leur somme donne au maximum 2. On conclut que $T = \frac{X_{(1)}+X_{(n)}}{2} \in (0,1)$.

Alors, si $0 \leq t \leq 1/2$, on doit avoir $0 < r < 2t$. Par contre, si $1/2 < t < 1$, il faut que $0 < r < 2(1-t)$. On obtient

$$f_T(t) = \begin{cases} n(2t)^{n-1}, & 0 < t < 1/2 \\ n(2(1-t))^{n-1}, & 1/2 < t < 1. \end{cases}$$

Ainsi,

$$\begin{aligned} E[T] &= 2^{n-1}n\left(\int_0^{1/2} t^n\, dt + \int_{1/2}^1 t(1-t)^{n-1}\, dt\right) \\ &= 2^{n-1}n\left(\frac{0{,}5^{n+1}}{n+1} + \int_{1/2}^1 t(1-t)^{n-1}\, dt\right) \end{aligned}$$

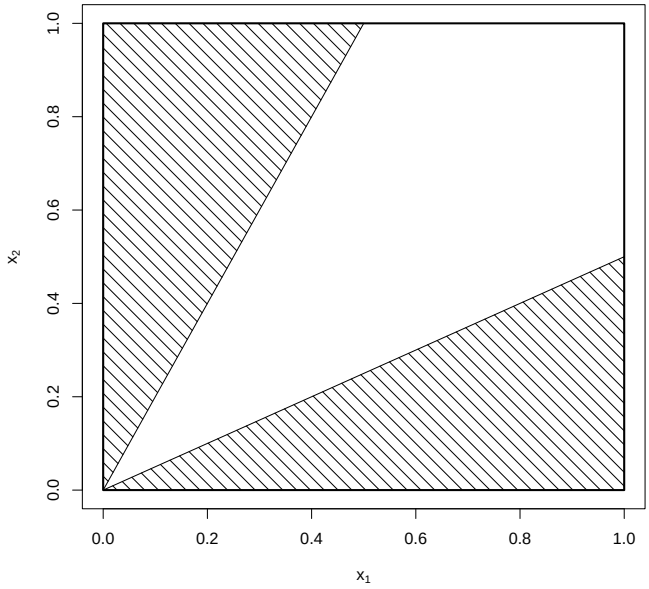


FIG. B.1 – Domaine de définition de $f_{X_1X_2}(x_1,x_2) = 4(1-x_1)(1-x_2)$, $x_1,x_2 \in (0,1)$. Les zones hachurées représentent les aires où $x_2 > 2x_1$ ou $x_1 > 2x_2$.

Pour l'intégrale entre $t = \frac{1}{2}$ et $t = 1$, on utiliser l'intégration par parties en posant $u = t$ et $\mathrm{d}v = (1-t)^{n-1}\mathrm{d}t$. Ainsi, on obtient

$$\begin{aligned} E[T] &= 2^{n-1}n\left(\frac{0{,}5^{n+1}}{n+1} + \left[\frac{t(1-t)^n}{n}\right]_{\frac{1}{2}}^{1} + \int_{1/2}^{1}\frac{(1-t)^n}{n}\,dt\right) \\ &= 2^{n-1}n\left(\frac{0{,}5^{n+1}}{n+1} + \frac{1}{2^{n+1}n} + \frac{1}{n(n+1)2^{n+1}}\right) \\ &= 2^{n-1}n\left(\frac{0{,}5^{n+1}}{n+1} + \frac{n+2}{2^{n+1}n(n+1)}\right) \\ &= \frac{1}{2} \end{aligned}$$

1.22 On sait que $X_i \sim U(0,1)$. On trouve la fonction de répartition du minimum $X_{(1)}$ avec

$$F_{X_{(1)}}(x) = 1 - (1 - F_X(x))^n = 1 - (1-x)^n,$$

d'où sa fonction de densité est

$$f_{X_{(1)}}(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_{X_{(1)}}(x) = n(1-x)^{n-1}$$

qui peut se réécrire sous la forme

$$f_{X_{(1)}}(x) = \frac{\Gamma(1+n)}{\Gamma(1)\Gamma(n)}x^{1-1}(1-x)^{n-1}.$$

On remarque que $X_{(1)} \sim \text{Bêta}(1,n)$. De la même façon pour $X_{(n)}$,

$$F_{X_{(n)}}(x) = (F_X(x))^n = x^n \text{ et } f_{X_{(n)}}(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_{X_{(n)}}(x) = nx^{n-1}$$

qui peut se réécrire sous la forme

$$f_{X_{(n)}}(x) = \frac{\Gamma(n+1)}{\Gamma(n)\Gamma(1)} x^{n-1}(1-x)^{1-1}.$$

On remarque que $X_{(n)} \sim \text{Bêta}(n,1)$. Ainsi, depuis l'exercice 1.16, la densité conjointe de $(X_{(1)},X_{(n)})$ peut être exprimée comme suit, pour $0 < x < y < 1$,

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)(y-x)^{n-2}.$$

Ainsi,

$$E\left[X_{(1)}X_{(n)}\right] = n(n-1)\int_0^1 x \int_x^1 y(y-x)^{n-2}\,dy\,dx.$$

L'intégrale intérieure ci-dessus se résoud par parties en posant $u = y$ et $dv = (y-x)^{n-2}\,dy$. On obtient alors

$$\begin{aligned} E\left[X_{(1)}X_{(n)}\right] &= n(n-1)\int_0^1 x\left(\frac{(y-x)^{n-1}y}{n-1} - \frac{1}{n-1}\int_x^1 (y-x)^{n-1}\,dy\right)dx \\ &= n\int_0^1 x(1-x)^{n-1}\,dx - \int_0^1 x(1-x)^n\,dx \\ &= \frac{1}{n+1} - \frac{1}{(n+1)(n+2)} \\ &= \frac{1}{n+2}, \end{aligned}$$

en intégrant une seconde fois par parties. Par conséquent,

$$\begin{aligned} \operatorname{cov}\left(X_{(1)},X_{(n)}\right) &= E\left[X_{(1)}X_{(n)}\right] - E\left[X_{(1)}\right]E\left[X_{(n)}\right] \\ &= \frac{1}{n+2} - \left(\frac{1}{n+1}\right)\left(\frac{n}{n+1}\right) \\ &= \frac{1}{(n+1)^2(n+2)}. \end{aligned}$$

a) On a

$$\begin{aligned} E[R] &= E\left[X_{(n)}\right] - E\left[X_{(1)}\right] \\ &= \frac{n}{n+1} - \frac{1}{n+1} \\ &= \frac{n-1}{n+1} \end{aligned}$$

et

$$\begin{aligned}\text{var}[R] &= \text{var}\big[X_{(1)}\big] + \text{var}\big[X_{(n)}\big] - 2\text{cov}\big(X_{(1)}, X_{(n)}\big) \\ &= \frac{n}{(n+1)^2(n+2)} + \frac{n}{(n+1)^2(n+2)} - \frac{2}{(n+1)^2(n+2)} \\ &= \frac{2n-2}{(n+1)^2(n+2)}.\end{aligned}$$

b) On a

$$\begin{aligned}E[T] &= \frac{E\big[X_{(1)}\big] + E\big[X_{(n)}\big]}{2} \\ &= \frac{1}{2}\left(\frac{n}{n+1} + \frac{1}{n+1}\right) \\ &= \frac{1}{2}\end{aligned}$$

et

$$\begin{aligned}\text{var}[T] &= \frac{\text{var}\big[X_{(1)}\big] + \text{var}\big[X_{(n)}\big] + 2\text{cov}\big(X_{(1)}, X_{(n)}\big)}{4} \\ &= \frac{1}{4}\left[\frac{n}{(n+1)^2(n+2)} + \frac{n}{(n+1)^2(n+2)} + \frac{2}{(n+1)^2(n+2)}\right] \\ &= \frac{1}{2(n+1)(n+2)}.\end{aligned}$$

Chapitre 2

2.1 On sait que $\bar{X}_5 \sim \mathcal{N}(E[\bar{X}_5], \text{var}[\bar{X}_5])$ avec $E[\bar{X}_5] = E[X] = 10$ et $\text{var}[\bar{X}_5] = \text{var}[X]/n = 125/5 = 5$. Par conséquent,

$$\begin{aligned}\Pr[\bar{X}_5 < c] &= \Pr\left[\frac{\bar{X}_5 - 10}{\sqrt{25}} < \frac{c-10}{\sqrt{25}}\right] \\ &= \Pr[Z < z_\alpha] \\ &= 1 - \alpha\end{aligned}$$

avec $Z \sim \mathcal{N}(0,1)$ et $z_\alpha = (c-10)/5$. Ici, on a $1-\alpha = 0{,}90$. On trouve dans une table de quantiles de la loi normale que $z_{0,10} = 1{,}282$, d'où $c = 16{,}41$.

2.2 On a $E[\bar{X}_n] = E[X] = \mu$, $\text{var}[\bar{X}_n] = \text{var}[X]/n = 100/n$ et $\bar{X}_n \sim \mathcal{N}(\mu, 100/n)$. Ainsi, on cherche n tel que

$$\begin{aligned}\Pr[\mu - 5 < \bar{X}_n < \mu + 5] &= \Pr\left[-\frac{5}{10/\sqrt{n}} < \frac{\bar{X}_n - \mu}{10/\sqrt{n}} < \frac{5}{10/\sqrt{n}}\right] \\ &= \Phi\left(\frac{5\sqrt{n}}{10}\right) - \Phi\left(-\frac{5\sqrt{n}}{10}\right) \\ &= 2\Phi\left(\frac{5\sqrt{n}}{10}\right) - 1 \\ &= 0{,}954,\end{aligned}$$

soit

$$\Phi\left(\frac{5\sqrt{n}}{10}\right) = 0{,}977.$$

On trouve dans une table de loi normale que $5\sqrt{n}/10 = 2$, d'où $n = 16$.

2.3 On a $\bar{X}_{25} \sim \mathcal{N}(0{,}16/25)$, $\bar{Y}_{25} \sim \mathcal{N}(1{,}9/25)$ et, par conséquent, $\bar{X}_{25} - \bar{Y}_{25} \sim \mathcal{N}(-1{,}1)$. On a donc

$$\begin{aligned}
\Pr[\bar{X}_{25} > \bar{Y}_{25}] &= \Pr[\bar{X}_{25} - \bar{Y}_{25} > 0] \\
&= \Pr\left[\frac{\bar{X}_{25} - \bar{Y}_{25} - (-1)}{\sqrt{1}} > \frac{0 - (-1)}{\sqrt{1}}\right] \\
&= 1 - \Phi(1) \\
&= 0{,}159.
\end{aligned}$$

2.4 a) La statistique $W = Y_1^2 + \cdots + Y_7^2$ suit une loi khi carré avec 7 degrés de liberté, $\chi^2_{(7)}$, puisque W est une somme de sept variables **indépendantes**, chacune d'entre elle étant le carré d'une variable aléatoire normale centrée réduite.

b) Si $Y_1, \ldots, Y_n$ forme un échantillon aléatoire tiré d'une $\mathcal{N}(\mu, \sigma^2)$, alors

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$$

a une distribution khi carré avec $n-1$ degrés de liberté, $\chi^2_{(n-1)}$. Si on pose $n = 7$ et $\sigma^2 = 1$, on trouve que

$$U = \sum_{i=1}^{7}(Y_i - \bar{Y})^2 \sim \chi^2_{(6)}.$$

c) Selon (b), U a une distribution $\chi^2_{(6)}$. Cela signifie que la somme a la même distribution que $Z_1^2 + \cdots + Z_6^2$, où $Z_1, \ldots, Z_6$ sont iid $\mathcal{N}(0{,}1)$ et indépendantes de Y_8, qui suit aussi une $\mathcal{N}(0{,}1)$. Ainsi,

$$Y_8^2 + U \sim \chi^2_{(7)}.$$

d) D'abord, on observe que $\sqrt{7}Y_8/\sqrt{W} = Y_8/\sqrt{W/7}$, où $Y_8 \sim \mathcal{N}(0{,}1)$ et $W \sim \chi^2_{(7)}$ selon a). De plus, W et Y_8 sont des variables aléatoires indépendantes. Ainsi, $Y_8/\sqrt{W/7}$ suit une loi Student $t_{(7)}$.

e) On note que $U \sim \chi^2_{(6)}$ selon b). Ensuite, U est indépendant de $Y_8 \sim \mathcal{N}(0{,}1)$. Ainsi,

$$\frac{\sqrt{6}Y_8}{\sqrt{U}} = \frac{Y_8}{\sqrt{U/6}} \sim t_{(6)}.$$

f) On sait que la moyenne échantillonnale $\bar{Y}$ est indépendante de la variance échantillonnale $U/6$. Donc, $\bar{Y}$ et Y_8 sont toutes deux indépendantes de U. Puisque $\bar{Y} \sim \mathcal{N}(0{,}1/7)$, on a $\sqrt{7}\bar{Y} \sim \mathcal{N}(0{,}1)$ et ainsi

$$7\bar{Y}^2 + Y_8^2 = (\sqrt{7}\bar{Y})^2 + Y_8^2 \sim \chi^2_{(2)}.$$

De plus, $U \sim \chi^2_{(6)}$ avec (b). Il en découle que

$$\frac{3(7\bar{Y}^2 + Y_8^2)}{U} = \frac{\frac{(\sqrt{7}\bar{Y})^2 + Y_8^2}{2}}{U/6} \sim \mathcal{F}_{(2,6)}.$$

2.5 On sait que $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$, que l'espérance d'une loi χ^2 est égale à son nombre de degrés de liberté et que sa variance est égale à deux fois son nombre de degrés de liberté. On a donc

$$\begin{aligned} E[S_n^2] &= \frac{\sigma^2}{(n-1)} E\left[\frac{(n-1)S_n^2}{\sigma^2}\right] \\ &= \frac{(n-1)\sigma^2}{(n-1)} = \sigma^2 \end{aligned}$$

et

$$\begin{aligned} \text{var}[S_n^2] &= \frac{\sigma^4}{(n-1)^2} \text{var}\left[\frac{(n-1)S_n^2}{\sigma^2}\right] \\ &= \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{(n-1)} \end{aligned}$$

2.6 On sait que $5S_6^2/\sigma^2 \sim \chi^2(5)$. Soit $Y \sim \chi^2(5)$. On a donc,

$$\begin{aligned} \Pr[2{,}30 < S_6^2 < 22{,}2] &= \Pr\left[\frac{5(2{,}30)}{10} < \frac{5S_6^2}{10} < \frac{5(22{,}2)}{10}\right] \\ &= \Pr[1{,}15 < Y < 11{,}1] \\ &= \Pr[Y < 11{,}1] - \Pr[Y < 1{,}15]. \end{aligned}$$

On trouve dans une table de quantiles de la loi khi carré (ou avec la fonction `qchisq` dans R, par exemple) que $\Pr[Y < 11{,}1] = 0{,}95$ et $\Pr[Y < 1{,}15] = 0{,}05$. Par conséquent, $\Pr[2{,}30 < S_6^2 < 22{,}2] = 0{,}90$.

2.7 On sait que $W = (n-1)S_n^2/\sigma^2 \sim \chi^2_{(n-1)}$, et on cherche à trouver la fonction de densité de $S_n^2 = \sigma^2 W/(n-1)$. Puisque c'est une transformation d'échelle, on peut utiliser la méthode de la fonction de répartition. D'abord, on exprime la fonction de répartition de S_n^2 en termes de la fonction de répartition de W :

$$F_{S_n^2}(s) = \Pr[S_n^2 \leq s] = \Pr\left[\frac{\sigma^2 W}{n-1} \leq s\right] = \Pr[W \leq s(n-1)/\sigma^2] = F_W\left(\frac{s(n-1)}{\sigma^2}\right).$$

Ensuite, on dérive $F_{S_n^2}(s)$ en fonction de s pour trouver la densité :

$$f_{S_n^2}(s) = \frac{\mathrm{d}}{\mathrm{d}s} F_W\left(\frac{s(n-1)}{\sigma^2}\right) = f_W\left(\frac{s(n-1)}{\sigma^2}\right)\frac{n-1}{\sigma^2}.$$

Finalement, la densité de W est la distribution d'une χ^2_{n-1}. Donc, pour $s > 0$,

$$f_{S_n^2}(s) = \frac{n-1}{\sigma^2} \frac{1}{2^{(n-1)/2}\Gamma(\frac{n-1}{2})} \left(\frac{(n-1)s}{\sigma^2}\right)^{(n-1)/2-1} \exp\left(-\frac{(n-1)s}{2\sigma^2}\right)$$
$$= \left(\frac{n-1}{2\sigma^2}\right)^{(n-1)/2} \frac{1}{\Gamma(\frac{n-1}{2})} s^{(n-1)/2-1} \exp\left(-\frac{s}{2\sigma^2/(n-1)}\right).$$

Par conséquent, S_n^2 suit une distribution gamma avec paramètres $\alpha = (n-1)/2$ et $\beta = 2\sigma^2/(n-1)$. La moyenne et la variance sont :

$$\mathrm{E}[S_n^2] = \alpha\beta = \frac{n-1}{2}\frac{2\sigma^2}{n-1} = \sigma^2$$
$$\mathrm{var}(S_n^2) = \alpha\beta^2 = \frac{n-1}{2}\frac{4\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1}.$$

2.8 On a $X = \mu + \sigma Z$. Par la technique de la fonction de répartition, on obtient :

$$F_X(x) = \Pr[X \le x]$$
$$= \Pr[\mu + \sigma Z \le x]$$
$$= \Pr\left[Z \le \frac{x-\mu}{\sigma}\right]$$
$$= \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Ainsi, la fonction de densité de probabilité de X est

$$f_X(x) = \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

2.9 Soit $Z \sim \mathcal{N}(0,1)$. La fonction génératrice des moments de la variable aléatoire Z^2 est

$$M_{Z^2}(t) = E\left[e^{Z^2 t}\right]$$
$$= \int_{-\infty}^{\infty} e^{z^2 t}\phi(z)\,dz$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{z^2 t} e^{-z^2/2}\,dz$$
$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2(1-2t)/2}\,dz.$$

En posant $\sigma^2 = (1-2t)^{-1}$, on voit que l'on peut écrire l'expression ci-dessus sous la forme

$$M_{Z^2}(t) = \sigma \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/(2\sigma^2)}\,dz.$$

On reconnaît alors sous l'intégrale la densité d'une loi normale de moyenne 0 et de variance σ^2. Par conséquent,

$$M_{Z^2}(t) = \sigma = \left(\frac{1}{1-2t}\right)^{1/2},$$

soit la fonction génératrice des moments d'une loi gamma de paramètres $\alpha = 1/2$ et $\beta = 2$ ou, de manière équivalente, d'une distribution $\chi^2(1)$.

2.10 a) On a $X \sim \mathcal{N}(0, \sigma^2)$ et $Y = X^2$. Il faut voir que $Y = X^2$ n'est pas une transformation bijective. On pose donc d'abord $Z = |X|$ et on trouve la densité de Z à l'aide de la technique de la fonction de répartition :

$$\begin{aligned} F_Z(z) &= \Pr[|X| \leq z] \\ &= \Pr[-z \leq X \leq z] \\ &= F_X(z) - F_X(-z) \end{aligned}$$

d'où

$$\begin{aligned} f_Z(z) &= f_X(z) + f_X(-z) \\ &= \frac{2}{\sigma\sqrt{2\pi}} e^{-x^2/(2\sigma^2)}, \quad z > 0. \end{aligned}$$

Ensuite, on pose la transformation bijective $Y = Z^2 = |X|^2 = X^2$. Par la technique du changement de variable, on a

$$\begin{aligned} f_Y(y) &= f_Z(y^{1/2}) \left| \frac{1}{2\sqrt{y}} \right| \\ &= \frac{2}{\sigma\sqrt{2\pi}} e^{-y/(2\sigma^2)} \left(\frac{1}{2\sqrt{y}} \right) \\ &= \frac{(2\sigma^2)^{-1/2}}{\Gamma(\frac{1}{2})} y^{1/2-1} e^{-y/(2\sigma^2)}, \quad y > 0 \end{aligned}$$

puisque $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. On voit donc que

$$Y \sim \text{Gamma}\left(\frac{1}{2}, 2\sigma^2\right).$$

Plus directement, on peut aussi voir que

$$X = \pm\sqrt{y}.$$

Par la méthode du changement de variable, on développe

$$\begin{aligned} F_Y(y) &= \Pr[Y \leq y] \\ &= \Pr[X^2 \leq y] \\ &= \Pr[-\sqrt{y} \leq X \leq \sqrt{y}] \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

On dérive ensuite pour trouver la fonction de densité :

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) &= \frac{1}{2\sqrt{y}}f_x(\sqrt{y}) - \frac{-1}{2\sqrt{y}}f_x(\sqrt{y}) \\ &= \frac{f_x(\sqrt{Y})}{\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi\sigma^2 y}}e^{-y\frac{1}{2\sigma^2}} \\ &= \frac{(2\sigma^2)^{-1/2}}{\Gamma(\frac{1}{2})}y^{1/2-1}e^{-y/(2\sigma^2)}, \quad y > 0.\end{aligned}$$

b) On sait que $Z = X_1 - X_2 \sim \mathcal{N}(0,2)$ et que $Z/\sqrt{2} \sim \mathcal{N}(0,1)$. En utilisant le résultat de la partie a), on a immédiatement que $Y = Z^2/2 \sim \chi^2(1)$.

2.11 a) Puisque la loi t est symétrique autour de zéro, on a

$$\begin{aligned}\Pr[|T| > 2{,}228] &= \Pr[T > 2{,}228] + \Pr[T < -2{,}228] \\ &= 2\Pr[T > 2{,}228].\end{aligned}$$

Or, on trouve dans la table de la loi t de l'annexe A.3 que $\Pr[T \leq 2{,}228] = 0{,}975$ si $T \sim t(10)$. Par conséquent, $\Pr[|T| > 2{,}228] = 2(1 - 0{,}975) = 0{,}05$.

b) Toutes les fonction R servant à évaluer des fonctions de répartition ont un argument `lower.tail`. Ce argument est `TRUE` par défaut, mais lorsque qu'il est `FALSE`, la fonction retourne la probabilité *au-dessus* du point `x`. Ainsi, la probabilité cherchée ici est

```
2 * pt(2.228, 10, lower.tail = FALSE)

## [1] 0.05001177
```

Il est recommandé d'utiliser cette approche parce qu'elle est, de manière générale, plus précise que le calcul du type `1 - pt(x, n)`, surtout loin dans les queues des distributions.

2.12 a) Par symétrie de la loi t,

$$\begin{aligned}\Pr[-b < T < b] &= \Pr[T < b] - \Pr[T < -b] \\ &= \Pr[T < b] - (1 - \Pr[T < b]) \\ &= 2\Pr[T < b] - 1 \\ &= 0{,}90.\end{aligned}$$

On cherche donc la valeur de b tel que $\Pr[T < b] = (1 + 0{,}90)/2 = 0{,}95$, où $T \sim t(14)$. Dans la table de la loi t de l'annexe A.3 on trouve que $b = 1{,}761$.

b) En définitive, on cherche le 95$^{\text{e}}$ centile d'une loi $t(14)$. Avec R, on obtient

```
qt(0.95, 14)

## [1] 1.76131
```

2.13 a) On a $U \sim \chi^2(r_1)$ et $V \sim \chi^2(r_2)$. On pose $F = (U/r_1)/(V/r_2)$ et, disons, $G = V$. Pour trouver la densité (marginale) de F, il faudra passer par la densité conjointe de F et G.

Les équations régissant la transformation de variables aléatoires sont

$$x = \frac{r_2}{r_1}\left(\frac{u}{v}\right) \qquad\qquad u = \frac{r_1}{r_2}xy$$
$$y = v \qquad\qquad v = y.$$

Ainsi, le jacobien de la transformation est

$$J = \begin{vmatrix} r_1 y/r_2 & r_1 x/r_2 \\ 0 & 1 \end{vmatrix} = \frac{r_1}{r_2} y$$

et la densité conjointe de F et G est

$$\begin{aligned} f_{FG}(x,y) &= f_{UV}\left(\frac{r_1}{r_2}xy, y\right)\left|\frac{r_1}{r_2}y\right| \\ &= \left(\frac{r_1}{r_2}\right) y f_U\left(\frac{r_1}{r_2}xy\right) f_V(y) \\ &= \frac{(r_1/r_2)(1/2)^{(r_1+r_2)/2}(r_1 xy/r_2)^{r_1/2-1} y^{r_2/2} e^{-(r_1 x/r_2+1)y/2}}{\Gamma(r_1/2)\Gamma(r_2/2)} \end{aligned}$$

pour $x > 0$ et $y > 0$. En intégrant, on trouve la densité marginale de F :

$$\begin{aligned} f_F(x) &= \int_0^\infty f_{FG}(x,y)\,dy \\ &= \frac{(r_1/r_2)^{(r_1+r_2)/2} x^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)} \\ &\quad \times \int_0^\infty \left(\frac{1}{2}\right)^{(r_1+r_2)/2} y^{(r_1+r_2)/2-1} e^{-(r_1 x/r_2+1)y/2}\,dy \\ &= \frac{\Gamma((r_1+r_2)/2)(r_1/r_2)^{r_1/2} x^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(r_1 x/r_2+1)^{(r_1+r_2)/2}} \\ &\quad \times \int_0^\infty \frac{(1/2)^{(r_1+r_2)/2}(r_1 x/r_2+1)^{(r_1+r_2)/2}}{\Gamma((r_1+r_2)/2)} y^{(r_1+r_2)/2-1} e^{-(r_1 x/r_2+1)y/2}\,dy \\ &= \frac{\Gamma((r_1+r_2)/2)(r_1/r_2)^{r_1/2} x^{r_1/2-1}}{\Gamma(r_1/2)\Gamma(r_2/2)(1+r_1 x/r_2)^{(r_1+r_2)/2}}, \end{aligned}$$

puisque l'intégrande ci-dessus est la densité d'une loi gamma. La loi de la variable aléatoire F est appelée loi F avec r_1 et r_2 degrés de liberté.

b) Par indépendance entre les variables aléatoires U et V, on a

$$\begin{aligned} E[F] &= E\left[\frac{U/r_1}{V/r_2}\right] \\ &= \frac{r_2}{r_1} E\left[\frac{U}{V}\right] \\ &= \frac{r_2}{r_1} E[U] E\left[\frac{1}{V}\right]. \end{aligned}$$

Or, $E[U] = r_1$ et

$$\begin{aligned} E\left[\frac{1}{V}\right] &= \int_0^\infty \frac{1}{v} f_V(v)\, dv \\ &= \int_0^\infty \frac{1}{2^{r_2/2}\Gamma(r_2/2)} v^{r_2/2-1-1} e^{-v/2}\, dv \\ &= \frac{2^{r_2/2-1}\Gamma(r_2/2-1)}{2^{r_2/2}\Gamma(r_2/2)}, \quad \frac{r_2}{2} - 1 > 0. \end{aligned}$$

Avec la propriété de la fonction gamma $\Gamma(x) = (x-1)\Gamma(x-1)$, cette expression se simplifie en

$$E\left[\frac{1}{V}\right] = \frac{1}{r_2 - 2},$$

d'où, enfin,

$$E[F] = \frac{r_2}{r_2 - 2}$$

pour $r_2 > 2$.

c) En procédant comme en b), on trouve que $E[U^2] = \text{var}[U] + E[U]^2 = 2r_1 + r_1^2$, que

$$E\left[\frac{1}{V^2}\right] = \frac{1}{(r_2-2)(r_2-4)}$$

et donc que

$$\begin{aligned} E[F^2] &= \frac{r_2^2}{r_1^2} E[U^2] E\left[\frac{1}{V^2}\right] \\ &= \frac{r_2^2(r_1+2)}{r_1(r_2-2)(r_2-4)}. \end{aligned}$$

Par conséquent,

$$\begin{aligned} \text{var}[F] &= \frac{r_2^2(r_1+2)}{r_1(r_2-2)(r_2-4)} - \left(\frac{r_2}{r_2-2}\right)^2 \\ &= 2\left(\frac{r_2}{r_2-2}\right)^2 \left(\frac{r_2+r_1-2}{r_1(r_2-4)}\right), \end{aligned}$$

pour $r_2 > 4$.

2.14 On a

$$\begin{aligned}\frac{1}{F} &= \left(\frac{U/\nu_1}{V/\nu_2}\right)^{-1} \\ &= \frac{V/\nu_2}{U/\nu_1}\end{aligned}$$

où $U \sim \chi^2(\nu_1)$ et $V \sim \chi^2(\nu_2)$. Puisqu'il s'agit d'un ratio de deux variables aléatoires χ^2 divisées chacune par son nombre de degrés de liberté, on a donc que

$$\frac{1}{F} \sim F(\nu_2, \nu_1).$$

2.15 On a $F \sim F(5,10)$ et l'on cherche a et b tel que $\Pr[F \leq a] = 0{,}05$ et $\Pr[F \leq b] = 0{,}95$. Dans une table de loi F, on trouve que $\Pr[F \leq 3{,}326] = 0{,}95$ et donc que $b = 3{,}33$. Puisque les quantiles inférieurs ne sont pas inclus dans la table de l'annexe A.4, on doit utiliser pour trouver a la relation $\Pr[F \leq a] = 1 - \Pr[F^{-1} \leq a^{-1}]$ où, tel que démontré à l'exercice **??**.14, $F^{-1} \sim F(10,5)$. Dans une table, on trouve que $a^{-1} = 4{,}74$, d'où $a = 0{,}211$.

Avec R, on obtient les mêmes résultats encore plus simplement :

```
qf(c(0.05, 0.95), 5, 10)

## [1] 0.2111904 3.3258345
```

2.16 On sait que $W^2 \sim \chi^2(1)$. Ainsi,

$$T^2 = \frac{W^2/1}{V/r},$$

qui est un ratio de deux variables aléatoires χ^2 divisées par leur nombre de degrés de liberté. Par définition de la loi F, on a donc que $T^2 \sim F(1,r)$.

2.17 Soit

$$Y = \sum_{i=1}^{\alpha} X_i$$

avec $X_i \sim$ Exponentielle(β) et $X_1, \ldots, X_\alpha$ indépendantes. Par le Théorème central limite,

$$\lim_{\alpha\to\infty} Y = \lim_{\alpha\to\infty} \sum_{i=1}^{\alpha} X_i \sim \mathcal{N}(\alpha E[X_i], \alpha \text{var}[X_i]).$$

Par conséquent,

$$\lim_{\alpha\to\infty} Y \sim N\left(\alpha\beta, \alpha\beta^2\right).$$

On trouve à la figure B.2 les graphiques de densités gamma pour quelques valeurs du paramètre α. On observe, en effet, que la distribution tend vers une normale lorsque α augmente.

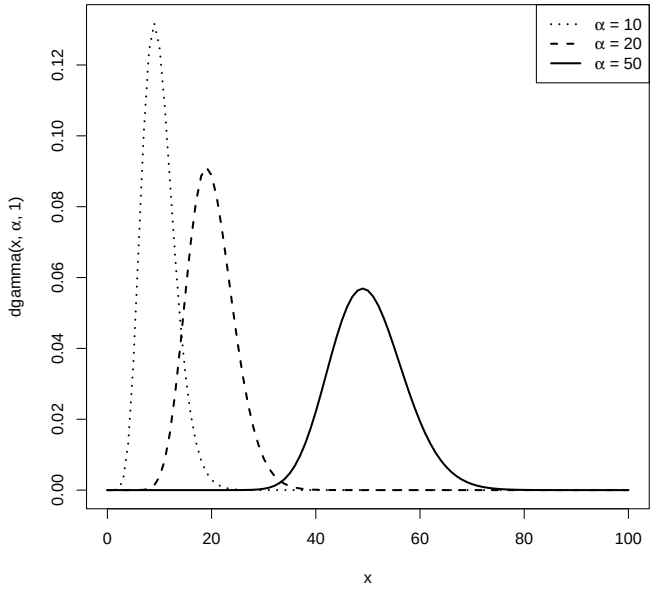


FIG. B.2 – Densités de lois gamma pour quelques valeurs du paramètre de forme α.

2.18 a) On a l'échantillon aléatoire $X_1, \ldots, X_{100}$, où

$$X_i \sim \chi^2(50) \equiv \text{Gamma}\left(\frac{50}{2}, 2\right), \quad i = 1, \ldots, 100.$$

Or, on sait que $Y = X_1 + \cdots + X_{100} \sim \text{Gamma}(100(25), 2)$ et que $\bar{X}_{100} = Y/100 \sim \text{Gamma}(2500, 2/100)$.

Ce résultat tient parce qu'une somme de k observations d'un échantillon aléatoire tiré d'une loi gamma est telle que $S = \sum_{i=1}^{k} X_i \sim Gamma(\alpha k, \beta)$. La distribution de $\bar{X}_{100}$ suit directement de la transformation d'échelle d'une loi gamma, qu'on peut démontrer en utilisant la fonction génératrice de moments. Ainsi,

$$\begin{aligned} \mathcal{M}_{\bar{X}_{100}}(t) &= E[e^{t\frac{Y}{100}}] \\ &= E[e^{Y\frac{t}{100}}] \\ &= \mathcal{M}_Y(\frac{t}{100}). \end{aligned}$$

b) On peut, par exemple, obtenir la probabilité demandée avec R ainsi :

```
pgamma(51, 2500, 50) - pgamma(49, 2500, 50)

## [1] 0.6827218
```

c) On a $E[\bar{X}_{100}] = 2500/50 = 50$ et $\text{var}[\bar{X}_{100}] = 2500/50^2 = 1$. En utilisant l'approximation normale, on trouve

$$\begin{aligned}\Pr[49 < \bar{X}_{100} < 51] &= \Pr\left[\frac{49-50}{1} < \frac{\bar{X}_{100}-50}{1} < \frac{51-50}{1}\right] \\ &\approx \Phi(1) - \Phi(-1) \\ &= 2\Phi(1) - 1 \\ &= 0{,}682.\end{aligned}$$

On trouve la valeur de $\Phi(1)$ dans une table de quantiles de la loi normale ou à l'aide d'un logiciel statistique.

2.19 Puique l'on ne demande qu'une valeur approximative pour $\Pr[7 < \bar{X} < 9]$, on va utiliser l'approximation normale. La taille de l'échantillon étant relativement grande, l'approximation sera très bonne. Soit $\bar{X} = (X_1 + \cdots + X_{128})/128$, où $X_i \sim \text{Gamma}(2,4)$, $i = 1,\ldots,128$. On a $E[\bar{X}] = E[X_i] = 8$ et $\text{var}[\bar{X}] = \text{var}[X_i]/128 = 1/4$. Par conséquent,

$$\begin{aligned}\Pr[7 < \bar{X} < 9] &= \Pr\left[\frac{7-8}{\sqrt{1/4}} < \frac{\bar{X}-8}{\sqrt{1/4}} < \frac{9-8}{\sqrt{1/4}}\right] \\ &\approx \Phi(2) - \Phi(-2) \\ &= 2\Phi(2) - 1 \\ &= 0{,}954.\end{aligned}$$

On trouve la valeur de $\Phi(2)$ dans une table de quantiles de la loi normale ou à l'aide d'un logiciel statistique.

2.20 On souhaitera utiliser l'approximation normale. Cela requiert de connaître les valeurs de l'espérance et de la variance de la moyenne de l'échantillon, $\bar{X}$, et, par ricochet, celles de la variable aléatoire avec densité $f(x)$. Or,

$$\begin{aligned}E[X] &= \int_0^1 3x^3\,dx = \frac{3}{4} \\ E[X^2] &= \int_0^1 3x^4\,dx = \frac{3}{5}\end{aligned}$$

et donc $\text{var}[X] = 3/80$. Ainsi, $E[\bar{X}] = E[X] = 3/4$ et $\text{var}[\bar{X}] = \text{var}[X]/15 = 1/400$ et

$$\begin{aligned}\Pr\left[\frac{3}{5} < \bar{X} < \frac{4}{5}\right] &= \Pr\left[\frac{3/5-3/4}{\sqrt{1/400}} < \frac{\bar{X}-3/4}{\sqrt{1/400}} < \frac{4/5-3/4}{\sqrt{1/400}}\right] \\ &\approx \Phi(1) - \Phi(-3) \\ &= 0{,}840.\end{aligned}$$

2.21 Pour chaque $i \in \{1,\ldots,n\}$, on pose $D_i = X_i - Y_i$. $D_1,\ldots,D_n$ sont mutuellement indépendants et identiquement distribués avec moyenne

$$\text{E}(D_1) = \text{E}(X_1) - \text{E}(Y_1) = \mu_1 - \mu_2$$

et variance

$$\text{var}(D_1) = \text{var}(X_1) + \text{var}(Y_1) = \sigma_1^2 + \sigma_2^2.$$

Par le Théorème central limite, on a donc que, quand $n \to \infty$,

$$\sqrt{n}\frac{\bar{D}_n - \text{E}[D_1]}{\sqrt{\text{var}(D_1)}} = \frac{(\bar{X}_n - \bar{Y}_n) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/n}} = Z_n$$

converge en distribution vers $\mathcal{N}(0,1)$.

2.22 a) La probabilité peut être calculée comme suit :

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| \leq 0{,}2) &= \Pr(-0{,}2 \leq \bar{X}_n - \mu \leq 0{,}2) \\ &= \Pr\left(-\frac{0{,}2}{\sqrt{2{,}5^2/9}} \leq \frac{\bar{X}_n - \mu}{\sqrt{2{,}5^2/9}} \leq \frac{0{,}2}{\sqrt{2{,}5^2/9}}\right).\end{aligned}$$

On sait que

$$\frac{\bar{X}_n - \mu}{\sqrt{2{,}5^2/9}} \sim \mathcal{N}(0,1),$$

la probabilité est alors

$$\begin{aligned}\Pr\left(-\frac{0{,}2}{\sqrt{2{,}5^2/9}} \leq \frac{\bar{X}_n - \mu}{\sqrt{2{,}5^2/9}} \leq \frac{0{,}2}{\sqrt{2{,}5^2/9}}\right) &= \Phi\left(\frac{0{,}2}{\sqrt{2{,}5^2/9}}\right) - \Phi\left(-\frac{0{,}2}{\sqrt{2{,}5^2/9}}\right) \\ &= 2\Phi\left(\frac{0{,}2}{\sqrt{2{,}5^2/9}}\right) - 1,\end{aligned}$$

où Φ représente la fonction de répartition de la loi normale centrée réduite. Cette probabilité peut être évaluée avec une table de la loi normale ou avec R ; ce dernier donne

```
2*pnorm(0.2/sqrt(2.5^2/9))-1
```

```
## [1] 0.1896697
```

b) De la même façon qu'en a), on trouve

$$\begin{aligned}\Pr(|\bar{X}_n - \mu| \leq 0{,}2) &= \Pr\left(-\frac{0{,}2}{\sqrt{2{,}5^2/n}} \leq \frac{\bar{X}_n - \mu}{\sqrt{2{,}5^2/9}} \leq \frac{0{,}2}{\sqrt{2{,}5^2/n}}\right) \\ &= 2\Phi\left(\frac{0{,}2}{\sqrt{2{,}5^2/n}}\right) - 1.\end{aligned}$$

On considère $n = 25$, $n = 36$ et $n = 64$ dans la formule, ce qui donne

```
n <- c(25,36,64)
2*pnorm(0.2/sqrt(2.5^2/n))-1
```

```
## [1] 0.3108435 0.3687726 0.4778274
```

On remarque que la probabilité que la moyenne échantillonnale diffère de la vraie moyenne d'au plus 0,2 once augmente avec n. C'est le résultat qu'on attend selon la Loi faible des grands nombres, qui indique que cette probabilité tend vers 1 quand $n \to \infty$.

c) Si

$$\Pr(|\bar{X}_n - \mu| \leq 0{,}2) = 2\Phi\left(\frac{0{,}2}{\sqrt{2{,}5^2/n}}\right) - 1 \geq 0{,}95,$$

alors $\Phi\left(0{,}2\sqrt{n}/2{,}5\right) \geq 0{,}975$ et ainsi $0{,}2\sqrt{n}/2{,}5 \geq 1{,}96$, ce qui implique que $\sqrt{n} \geq 24{,}5$ ou $n \geq 600{,}25$. La taille d'échantillon doit donc être de $n = 601$ pour que la probabilité soit *le plus près possible, mais pas plus petite que* 0,95.

d) Dans ce cas, la probabilité à calculer est

$$\Pr\left(-\frac{0{,}2}{\sqrt{s_n^2/9}} \leq \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/9}} \leq \frac{0{,}2}{\sqrt{s_n^2/9}}\right).$$

On sait que

$$\frac{\bar{X}_n - \mu}{\sqrt{S_n^2/9}} \sim t_{(8)},$$

on trouve que

$$\begin{aligned}\Pr\left(-\frac{0{,}2}{\sqrt{s_n^2/9}} \leq \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/9}} \leq \frac{0{,}2}{\sqrt{s_n^2/9}}\right) &= \Pr\left(-\frac{0{,}2}{\sqrt{5{,}5/9}} \leq \frac{\bar{X}_n - \mu}{\sqrt{5{,}5/9}} \leq \frac{0{,}2}{\sqrt{5{,}5/9}}\right) \\ &= T_{(8)}\left(\frac{0{,}2}{\sqrt{5{,}5/9}}\right) - T_{(8)}\left(-\frac{0{,}2}{\sqrt{5{,}5/9}}\right),\end{aligned}$$

où $T_{(8)}$ est la fonction de répartition d'une loi Student t avec 8 degrés de liberté. Encore une fois, cette expression peut être évaluée avec une table de la loi normale ou avec R ; ce dernier donne

```
pt(0.2/sqrt(5.5/9),df=8) -pt(-0.2/sqrt(5.5/9),df=8)

## [1] 0.1954708
```

La probabilité obtenue est plus grande qu'en a) étant donné la variabilité ajoutée avec l'estimation de la variance.

2.23 a) On peut utiliser le fait que

$$\frac{S_n^2/\sigma_1^2}{S_m^2/\sigma_2^2} \sim F_{(n-1,m-1)},$$

où σ_1^2 et σ_2^2 représentent la variance du premier et du second échantillon, respectivement. Dans ce contexte, $n = 10$, $m = 6$ et $\sigma_1^2 = 2\sigma_2^2$. Donc,

$$\frac{S_n^2}{2S_m^2} \sim F(9{,}5).$$

Si W représente une variable aléatoire $F_{(9,5)}$, on a

$$\Pr\left(\frac{S_n^2}{S_m^2} \leq b\right) = \Pr(W \leq b/2)$$

et donc $b/2$ est le 95e quantile de la distribution $F_{(9,5)}$.

```
qf(0.95,df1=9,df2=5)
```

```
## [1] 4.772466
```

Par conséquent,

$$b = 2 \times 4.7724656 = 9.5449312.$$

Pour trouver a, on note que

$$\frac{S_m^2/\sigma_2^2}{S_n^2/\sigma_1^2} = 2\frac{S_m^2}{S_n^2} \sim F_{(5,9)}.$$

Alors,

$$\Pr\left(\frac{S_n^2}{S_m^2} \geq a\right) = \Pr\left(\frac{S_m^2}{S_n^2} \leq \frac{1}{a}\right) = \Pr\left(2\frac{S_m^2}{S_n^2} \leq \frac{2}{a}\right).$$

Donc, $2/a$ est le 95e quantile de la distribution $F_{(5,9)}$.

```
qf(0.95,df1=5,df2=9)
```

```
## [1] 3.481659
```

et ainsi

$$a = \frac{2}{3.4816587} = 0.5744389.$$

b) Cette probabilité égale

$$\begin{aligned}\Pr\left(a \leq \frac{S_n^2}{S_m^2} \leq b\right) &= \Pr\left(\frac{S_n^2}{S_m^2} \leq b\right) - \Pr\left(\frac{S_n^2}{S_m^2} \leq a\right) \\ &= \Pr\left(\frac{S_n^2}{S_m^2} \leq b\right) + \Pr\left(\frac{S_n^2}{S_m^2} \geq a\right) - 1 = 2 \times 0{,}95 - 1 = 0{,}9.\end{aligned}$$

Chapitre 3

3.1 On a

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right] &= \frac{1}{n}\sum_{i=1}^{n}E[(X_i-\mu)^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}E[X_i^2-2\mu X_i+\mu^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}(E[X_i^2]-\mu^2) \\ &= \frac{1}{n}\sum_{i=1}^{n}[(\sigma^2+\mu^2)-\mu^2] \\ &= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 \\ &= \sigma^2, \end{aligned}$$

d'où l'expression du côté gauche de l'égalité est un estimateur sans biais du paramètre σ^2.

3.2 On a,

$$\begin{aligned} E[a_1X_1+\cdots+a_nX_n] &= E[a_1X_1]+\cdots+E[a_nX_n] \\ &= (a_1+\cdots+a_n)E[X_1] \\ &= (a_1+\cdots+a_n)\mu. \end{aligned}$$

Pour que $a_1X_1+\cdots+a_nX_n$ soit un estimateur sans biais de μ, il faut que $\sum_{i=1}^{n}a_i=1$.

3.3 a) Il faut d'abord calculer l'espérance de l'estimateur :

$$\begin{aligned} E[\bar{X}_n^2] &= \text{var}[\bar{X}_n]+E[\bar{X}_n]^2 \\ &= \frac{\sigma^2}{n}+\mu^2. \end{aligned}$$

On voit que $\bar{X}_n^2$ est un estimateur biaisé de μ^2 et que le biais est σ^2/n.

b) Puisque

$$\lim_{n\to\infty}E[\bar{X}_n^2]=\lim_{n\to\infty}\frac{\sigma^2}{n}+\mu^2=\mu^2,$$

$\bar{X}_n^2$ est un estimateur asymptotiquement sans biais de μ^2.

3.4 La fonction de densité de probabilité de la k^{e} statistique d'ordre est donnée par le Théorème 6.5 des notes de cours.

$$f_{X_{(k)}}(x)=\frac{n!}{(k-1)!(n-k)!}F(x)^{k-1}\{1-F(x)\}^{n-k}f(x)$$

Pour $n = 3$ et $X_i \sim U(0,\theta)$, $i = 1,2,3$, on a

$$\begin{aligned} f_{X_{(1)}}(x) &= 3\left(1 - \frac{x}{\theta}\right)^2 \left(\frac{1}{\theta}\right) \\ &= 3\left(\frac{1}{\theta} - \frac{2x}{\theta^2} + \frac{x^2}{\theta^3}\right), \quad 0 < x < \theta \end{aligned}$$

et

$$\begin{aligned} f_{X_{(2)}}(x) &= 6\left(\frac{x}{\theta}\right)\left(1 - \frac{x}{\theta}\right)\left(\frac{1}{\theta}\right) \\ &= 6\left(\frac{x}{\theta^2} - \frac{x^2}{\theta^3}\right), \quad 0 < x < \theta. \end{aligned}$$

Ainsi, d'une part,

$$\begin{aligned} E[4X_{(1)}] &= 4\int_0^\theta x f_{X_{(1)}}(x)\,dx \\ &= 12\int_0^\theta \left(\frac{x}{\theta} - \frac{2x^2}{\theta^2} + \frac{x^3}{\theta^3}\right) dx \\ &= 12\left(\frac{\theta}{2} - \frac{2\theta}{3} + \frac{\theta}{4}\right) \\ &= \theta \end{aligned}$$

et

$$\begin{aligned} E[2X_{(2)}] &= 2\int_0^\theta x f_{X_{(2)}}(x)\,dx \\ &= 12\int_0^\theta \left(\frac{x^2}{\theta^2} - \frac{x^3}{\theta^3}\right) dx \\ &= 12\left(\frac{\theta}{3} - \frac{\theta}{4}\right) \\ &= \theta, \end{aligned}$$

d'où $4X_{(1)}$ et $2X_{(2)}$ sont des estimateurs sans biais de θ. D'autre part,

$$\begin{aligned} E[(4X_{(1)})^2] &= 16\int_0^\theta x^2 f_{X_{(1)}}(x)\,dx \\ &= 48\int_0^\theta \left(\frac{x^2}{\theta} - \frac{2x^3}{\theta^2} + \frac{x^4}{\theta^3}\right) dx \\ &= 48\left(\frac{\theta^2}{3} - \frac{2\theta^2}{4} + \frac{\theta^2}{5}\right) \\ &= \frac{8\theta^2}{5} \end{aligned}$$

et

$$\begin{aligned} E[(2X_{(2)})^2] &= 4\int_0^\theta x^2 f_{X_{(2)}}(x)\,dx \\ &= 24\int_0^\theta \left(\frac{x^3}{\theta^3} - \frac{x^4}{\theta^3}\right) dx \\ &= 24\left(\frac{\theta^2}{4} - \frac{\theta^2}{5}\right) \\ &= \frac{6\theta^2}{5}. \end{aligned}$$

Par conséquent,

$$\text{var}[4X_{(1)}] = \frac{8\theta^2}{5} - \theta^2 = \frac{3\theta^2}{5}$$

et

$$\text{var}[2X_{(2)}] = \frac{6\theta^2}{5} - \theta^2 = \frac{\theta^2}{5}.$$

3.5 Soit $X_{(n)} = \max(X_1,\ldots,X_n)$ et $X_{(1)} = \min(X_1,\ldots,X_n)$, où $X_i \sim U(0,\theta)$, $i = 1,\ldots,n$. On sait alors que

$$\begin{aligned} f_{X_{(n)}}(x) &= n(F(x))^{n-1} f(x) \\ &= \frac{nx^{n-1}}{\theta^n}, \quad 0 < x < \theta \end{aligned}$$

et que

$$\begin{aligned} f_{X_{(1)}}(x) &= n(1 - F(x))^{n-1} f(x) \\ &= \frac{n(\theta - x)^{n-1}}{\theta^n}, \quad 0 < x < \theta. \end{aligned}$$

a) On souhaite développer un estimateur sans biais de θ basé sur $X_{(n)}$. En premier lieu,

$$\begin{aligned} E[X_{(n)}] &= \int_0^\theta x f_{X_{(n)}}(x)\,dx \\ &= \frac{n}{\theta^n}\int_0^\theta x^n\,dx \\ &= \frac{n\theta}{n+1}. \end{aligned}$$

Par définition, comme un estimateur sans biais $\hat{\theta}$ doit être trouvé, cet estimateur doit être tel que $E[\hat{\theta}] = \theta$. Il faut corriger $E[X_{(n)}]$ en inversant la fonction $g(\theta) = \frac{n\theta}{n+1}$. On remarque ainsi que

$$E[X_{(n)}] = \frac{n\theta}{n+1} \Leftrightarrow \theta = E[X_{(n)}]\frac{n+1}{n} = E\left[\frac{X_{(n)}(n+1)}{n}\right]$$

Un estimateur sans biais de θ est donc $\hat{\theta} = \frac{(n+1)X_{(n)}}{n}$.

b) Comme en a), on calcule d'abord l'espérance de la statistique :

$$\begin{aligned} E[X_{(1)}] &= \int_0^\theta x f_{X_{(1)}}(x)\,dx \\ &= \frac{n}{\theta^n}\int_0^\theta x(\theta - x)^{n-1} x\,dx \\ &= \frac{\theta}{n+1} \end{aligned}$$

en intégrant par parties. En corrigeant par la même méthode qu'en a), un estimateur sans biais de θ basé sur le minimum de l'échantillon est donc $\hat{\theta} = (n+1)X_{(1)}$.

3.6 On sait que $E[X] = np$ et que $\text{var}[X] = np(1-p)$. Or,

$$\begin{aligned} E\left[n\left(\frac{X}{n}\right)\left(1 - \frac{X}{n}\right)\right] &= E[X] - \frac{E[X^2]}{n} \\ &= np - \frac{\text{var}[X] + E[X]^2}{n} \\ &= np - \frac{np(1-p) + n^2p^2}{n} \\ &= (n-1)p(1-p) \\ &= np(1-p) - p(1-p). \end{aligned}$$

La statistique est donc un estimateur biaisé de la variance et le biais est $-p(1-p)$. La statistique sur-estime la variance.

3.7 Il faut démontrer que $\lim_{n\to\infty} \Pr[|X_{(1)} - \theta| < \epsilon] = 1$. On sait que si $X \sim U(\theta, \theta+1)$, alors

$$\begin{aligned} f_X(x) &= 1, \quad \theta < x < \theta + 1 \\ F_X(x) &= x - \theta, \quad \theta < x < \theta + 1 \end{aligned}$$

et

$$\begin{aligned} f_{X_{(1)}}(x) &= n f_X(x)(1 - F_X(x))^{n-1} \\ &= n(1 - x + \theta)^{n-1}, \quad \theta < x < \theta + 1. \end{aligned}$$

Ainsi,

$$\begin{aligned} \Pr[|X_{(1)} - \theta| < \epsilon] &= \Pr[\theta - \epsilon < X_{(1)} < \theta + \epsilon] \\ &= \int_\theta^{\theta+\epsilon} n(1 - x + \theta)^{n-1}\,dx \\ &= 1 - (1-\epsilon)^n. \end{aligned}$$

Or, cette dernière expression tend vers 1 lorsque n tend vers l'infini, ce qui complète la démonstration.

3.8 Puisque $\bar{X}$ est un estimateur sans biais de la moyenne d'une distribution, quelqu'elle fut, et que $\lim_{n\to\infty} \text{var}[\bar{X}] = \lim_{n\to\infty} \text{var}[X]/n = 0$, alors $\bar{X}$ est toujours un estimateur convergent de la moyenne.

3.9 a) On peut conclure que $\hat{\mu}_1$ est un estimateur sans biais de μ par le fait que

$$\text{E}(\hat{\mu}_1) = \frac{1}{2}\{\text{E}(X_1) + \text{E}(X_2)\} = \frac{1}{2}(\mu + \mu) = \mu.$$

De même, $\hat{\mu}_2$ est un estimateur sans biais de μ, car

$$\begin{aligned}
\text{E}(\hat{\mu}_2) &= \frac{1}{4}\,\text{E}(X_1) + \frac{1}{2(n-2)}\sum_{i=2}^{n-1}\text{E}(X_i) + \frac{1}{4}\,\text{E}(X_n) \\
&= \frac{1}{4}\,\mu + \frac{1}{2(n-2)}\,(n-2)\mu + \frac{1}{4}\,\mu \\
&= \frac{1}{4}\,\mu + \frac{1}{2}\,\mu + \frac{1}{4}\,\mu = \mu.
\end{aligned}$$

Finalement, $\hat{\mu}_3$ est un estimateur sans biais de μ puisque

$$\text{E}(\hat{\mu}_3) = \frac{1}{n}\sum_{i=1}^{n}\text{E}(X_i) = \frac{n\mu}{n} = \mu.$$

b) Par définition,

$$\text{eff}(\hat{\mu}_3, \hat{\mu}_2) = \frac{\text{var}(\hat{\mu}_2)}{\text{var}(\hat{\mu}_3)} \quad \text{et} \quad \text{eff}(\hat{\mu}_3, \hat{\mu}_1) = \frac{\text{var}(\hat{\mu}_1)}{\text{var}(\hat{\mu}_3)}.$$

Pour calculer ces ratios, on doit commencer par déterminer la variance de chacun des trois estimateurs. D'abord, on trouve

$$\text{var}(\hat{\mu}_1) = \frac{1}{4}\{\text{var}(X_1) + \text{var}(X_2)\} = \frac{\sigma^2}{2}.$$

Ensuite,

$$\begin{aligned}
\text{var}(\hat{\mu}_2) &= \frac{1}{16}\,\text{var}(X_1) + \frac{1}{4(n-2)^2}\sum_{i=2}^{n-1}\text{var}(X_i) + \frac{1}{16}\,\text{var}(X_n) \\
&= \frac{1}{16}\,\sigma^2 + \frac{1}{4(n-2)^2}\,(n-2)\sigma^2 + \frac{1}{16}\,\sigma^2 \\
&= \frac{2(n-2)+4}{16(n-2)}\,\sigma^2 \\
&= \frac{n}{8(n-2)}\,\sigma^2.
\end{aligned}$$

Finalement,

$$\text{var}(\hat{\mu}_3) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{n}{n^2}\,\sigma^2 = \frac{\sigma^2}{n}.$$

Il en découle que

$$\text{eff}(\hat{\mu}_3, \hat{\mu}_2) = \frac{\frac{n\sigma^2}{8(n-2)}}{\frac{\sigma^2}{n}} = \frac{n^2}{8(n-2)}$$

et

$$\text{eff}(\hat{\mu}_3, \hat{\mu}_1) = \frac{\frac{\sigma^2}{2}}{\frac{\sigma^2}{n}} = \frac{n}{2}.$$

c) Pour toute valeur de $n \geq 3$, on a

$$\text{eff}(\hat{\mu}_3, \hat{\mu}_1) > 1 \quad \text{et} \quad \text{eff}(\hat{\mu}_3, \hat{\mu}_2) > 1.$$

Donc, $\hat{\mu}_3$ est toujours préférable à $\hat{\mu}_1$ ou $\hat{\mu}_2$.

3.10 a) La fonction de répartition de $\hat{\beta} = \min(X_1, \ldots, X_n)$, pour $x \geq \beta$, est

$$\begin{aligned}
\Pr[\min(X_1, \ldots, X_n) \leq x] &= 1 - \Pr[\min(X_1, \ldots, X_n) > x] \\
&= 1 - \Pr[X_1 > x, \ldots, X_n > x] \\
&= 1 - \Pr[X_1 > x] \times \cdots \times \Pr[X_n > x], \text{ par indépendance} \\
&= 1 - \{\Pr[X_1 > x]\}^n, \text{ par i.d.} \\
&= 1 - \left\{\frac{\beta}{x}\right\}^{\alpha n}
\end{aligned}$$

Donc, $\min(X_1, \ldots, X_n)$ a la même fonction de répartition que X avec un nouveau paramètre $\alpha^\star = \alpha n$. Ce qui signifie que la densité est

$$f_{\min}(x) = \alpha n \beta^{\alpha n} x^{-(\alpha n + 1)}, \quad x \geq \beta.$$

b) On utilise la définition d'un estimateur convergent. Si $\varepsilon > 0$,

$$\begin{aligned}
\Pr[|\hat{\beta} - \beta| \leq \varepsilon] &= \Pr[\beta - \varepsilon < \hat{\beta} \leq \beta + \varepsilon] \\
&= \Pr[\hat{\beta} \leq \beta + \varepsilon],
\end{aligned}$$

puisque $\Pr[\hat{\beta} < \beta] = 0$ par le fait que le domaine du minimum est le même que celui des observations, $[\beta, \infty)$. Ainsi, on remplace dans la fonction de répartition développée en a) :

$$\Pr[|\hat{\beta} - \beta| \leq \varepsilon] = \Pr[\hat{\beta} \leq \beta + \varepsilon] = F_{\min}(\beta + \varepsilon) = 1 - \left(\frac{\beta}{\beta + \varepsilon}\right)^{\alpha n}.$$

Puisque ε est positif, le ratio $\frac{\beta}{\beta+\varepsilon} < 1$ et donc $\left(\frac{\beta}{\beta+\varepsilon}\right)^{\alpha n} \to 0$ quand $n \to \infty$. Ainsi,

$$\lim_{n \to \infty} \Pr[|\hat{\beta} - \beta| \leq \varepsilon] = 1,$$

et $\hat{\beta}$ est convergent pour β.

c) On trouve d'abord $\mathrm{E}[\hat{\beta}]$ et $\mathrm{E}[\hat{\beta}^2]$:

$$\mathrm{E}[\hat{\beta}] = \int_\beta^\infty x\alpha n\beta^{\alpha n} x^{-(\alpha n+1)}\mathrm{d}x = \int_\beta^\infty \alpha n\beta^{\alpha n} x^{-\alpha n}\mathrm{d}x = \frac{\alpha n\beta}{\alpha n-1}$$
$$\mathrm{E}[\hat{\beta}^2] = \int_\beta^\infty x^2\alpha n\beta^{\alpha n} x^{-(\alpha n+1)}\mathrm{d}x = \int_\beta^\infty \alpha n\beta^{\alpha n} x^{-(\alpha n-1)}\mathrm{d}x = \frac{\alpha n\beta^2}{\alpha n-2}.$$

Le biais est donc

$$B(\hat{\beta}) = \mathrm{E}[\hat{\beta}] - \beta = \frac{\alpha n\beta}{\alpha n-1} - \beta = \frac{\beta}{\alpha n-1} \to 0$$

quand $n \to \infty$. L'estimateur est donc asymptotiquement sans biais.
La variance est

$$\mathrm{var}(\hat{\beta}) = \frac{\alpha n\beta^2}{\alpha n-2} - \left(\frac{\alpha n\beta}{\alpha n-1}\right)^2 = \frac{\alpha n\beta^2}{(\alpha n-2)(\alpha n-1)^2}$$

et l'erreur quadratique moyenne est

$$\mathrm{EQM}(\hat{\beta}) = \mathrm{var}(\hat{\beta}) + B^2(\hat{\beta}) = \frac{\alpha n\beta^2}{(\alpha n-2)(\alpha n-1)^2} + \frac{\beta^2}{(\alpha n-1)^2} = \frac{2\beta^2}{(\alpha n-1)(\alpha n-2)}.$$

d) Si α est connu, le paramètre inconnu est β et

$$f(x_1|\alpha,\beta) \times \cdots \times f(x_n|\alpha,\beta) = \underbrace{\left(\prod_{i=1}^n x_i\right)^{-(\alpha+1)}}_{h(x_1,\ldots,x_n)} \underbrace{(\alpha\beta^\alpha)^n \mathbf{1}\{\min(x_1,\ldots,x_n) \geq \beta\}}_{g\{\min(x_1,\ldots,x_n),\beta\}}.$$

Par le théorème de factorisation de Fisher–Neyman, $\min(X_1,\ldots,X_n)$ est une statistique exhaustive pour β.

e) On observe que

$$\begin{aligned} f(y_1|\alpha,\beta) \times \cdots \times f(x_n|\alpha,\beta) &= \alpha\beta^\alpha x_1^{-(\alpha+1)}\mathbf{1}(x_1 \geq \beta) \times \cdots \times \alpha\beta^\alpha x_n^{-(\alpha+1)}\mathbf{1}(x_n \geq \beta) \\ &= (\alpha\beta^\alpha)^n \left(\prod_{i=1}^n x_i\right)^{-(\alpha+1)} \times \mathbf{1}\{\min(x_1,\ldots,x_n) \geq \beta\}. \end{aligned}$$

Si β est connu, le paramètre incconu est α. Puisque

$$f(x_1|\alpha,\beta) \times \cdots \times f(x_n|\alpha,\beta) = \underbrace{(\alpha\beta^\alpha)^n \left(\prod_{i=1}^n x_i\right)^{-(\alpha+1)}}_{g(\prod x_i,\alpha)} \underbrace{\mathbf{1}\{\min(x_1,\ldots,x_n) \geq \beta\}}_{h(x_1,\ldots,x_n)},$$

on peut conclure par le théorème de factorisation de Fisher–Neyman que $X_1 \times \cdots \times X_n$ est une statistique exhaustive pour α.

f) Lorsque α et β sont inconnus,

$$f(x_1|\alpha,\beta)\times\cdots\times f(x_n|\alpha,\beta) = \underbrace{(\alpha\beta^\alpha)^n\left(\prod_{i=1}^n x_i\right)^{-(\alpha+1)}\mathbf{1}\{\min(x_1,\ldots,x_n)\geq\beta\}}_{g\{\prod x_i,\min(x_1,\ldots,x_n),\alpha,\beta\}}$$

donc les statistiques $X_1\times\cdots\times X_n$ et $\min(X_1,\ldots,X_n)$ sont conjointement exhaustives pour α et β.

3.11 On a un échantillon aléatoire de taille 1. Or,

$$\begin{aligned} f(x_1;\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-x_1^2/(2\sigma^2)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-|x_1|^2/(2\sigma^2)} \\ &= g(|x_1|;\sigma^2)h(x_1), \end{aligned}$$

avec

$$g(x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\,e^{-x^2/(2\sigma^2)}$$

et $h(x)=1$. Ainsi, par le théorème de factorisation de Fisher–Neyman, $|X_1|$ est une statistique exhaustive pour σ^2.

3.12 Soit $X_1,\ldots,X_n$ un échantillon aléatoire d'une distribution uniforme sur l'intervalle $(-\theta,\theta)$. La fonction de vraisemblance de cet échantillon est

$$f(x_1,\ldots,x_n;\theta) = \begin{cases}(2\theta)^{-n}, & -\theta<x_i<\theta, i=1,\ldots,n\\ 0, & \text{ailleurs.}\end{cases}$$

La fonction de vraisemblance est donc non nulle seulement si toutes les valeurs de l'échantillon se trouvent dans l'intervalle $(-\theta,\theta)$ ou, de manière équivalente, si

$$\begin{aligned} -\theta<x_i<\theta\ i\in 1,\ldots,n \\ -\theta<\min(x_1,\ldots,x_n)\text{ et }\max(x_1,\ldots,x_n)<\theta \\ \theta>-\min(x_1,\ldots,x_n)\text{ et }\max(x_1,\ldots,x_n)<\theta \\ \max_{i=1,\ldots,n}(|x_i|)<\theta. \end{aligned}$$

On peut donc, par exemple, réécrire la fonction de vraisemblance sous la forme

$$\begin{aligned} f(x_1,\ldots,x_n;\theta) &= \left(\frac{1}{2\theta}\right)^n \mathbf{1}_{\{0\leq\max_{i=1,\ldots,n}(|x_i|)\leq\theta)\}} \\ &= g\left(\max_{i=1,\ldots,n}(|x_i|);\theta\right)h(x_1,\ldots,x_n), \end{aligned}$$

avec

$$g(x;\theta) = \left(\frac{1}{2\theta}\right)^n I_{\{0\leq x\leq\theta)\}}$$

et $h(x_1,\ldots,x_n) = 1$. Ainsi, par le théorème de factorisation de Fisher–Neyman, on établit que $T = \max_{i=1,\ldots,n}(|X_i|)$ est une statistique exhaustive pour le paramètre θ. Une autre factorisation possible serait

$$f(x_1,\ldots,x_n;\theta) = \left(\frac{1}{2\theta}\right)^n I_{\{-\infty<x_{(n)}<\theta)\}} I_{\{-\theta<x_{(1)}<\infty)\}},$$

ce qui donne comme statistique exhaustive $T = (X_{(1)}, X_{(n)})$.

3.13 Nous allons démontrer un résultat général applicable à plusieurs distributions, dont la Poisson. Il existe une famille de distributions que l'on nomme la *famille exponentielle* (il ne s'agit pas d'une référence à la densité exponentielle, bien que cette dernière soit un cas particulier de la famille exponentielle). Cette famille comprend toutes les distributions dont la densité peut s'écrire sous la forme

$$f(x;\theta) = h(x)c(\theta)e^{\eta(\theta)t(x)},$$

où h, c, η et t sont des fonctions quelconques. Par exemple, la fonction de masse de probabilité de la loi de Poisson peut s'écrire comme suit :

$$\begin{aligned} f(x;\theta) &= \frac{\theta^x e^{-\theta}}{x!} \\ &= \left(\frac{1}{x!}\right) e^{-\theta} e^{\ln(\theta)x} \\ &= h(x)c(\theta)e^{\eta(\theta)t(x)}, \end{aligned}$$

avec $h(x) = (x!)^{-1}$, $c(\theta) = e^{-\theta}$, $\eta(\theta) = \ln\theta$ et $t(x) = x$. La loi est donc membre de la famille exponentielle. Les lois binomiale, gamma, normale et bêta font aussi partie de la famille exponentielle. En revanche, des lois comme l'uniforme sur (0,θ) et l'exponentielle translatée n'en font pas partie.

Pour tous les membres de la famille exponentielle, on a

$$f(x_1,\ldots,x_n;\theta) = \left(\prod_{i=1}^n h(x_i)\right)(c(\theta))^n e^{\eta(\theta)\sum_{i=1}^n x_i}.$$

Ainsi, on voit que le théorème de factorisation permet de conclure que la statistique

$$T(X_1,\ldots,X_n) = \sum_{i=1}^n X_i$$

est une statistique exhaustive pour le paramètre θ pour tous les membres de la famille exponentielle, dont la loi de Poisson.

3.14 a) On trouve que $\tilde{\lambda}$ est sans biais pour λ par le fait que

$$\mathrm{E}[\tilde{\lambda}] = \mathrm{E}[X_1] = \lambda.$$

b) Par le théorème de Rao–Blackwell, on trouve

$$\lambda^* = \mathrm{E}\left[\tilde{\lambda}|T\right] = \mathrm{E}\left[X_1|\sum_{i=1}^{n} X_i = t\right].$$

Par contre, étant donné que la somme des espérances est équivalente à l'espérance de la somme, on remarque que la sommation des espérances conditionnelles sur toutes les observations résulte en un résultat trivial sachant que $\sum_{i=1}^{n} X_i = t$:

$$\sum_{i=1}^{n} \mathrm{E}\left[X_i|\sum_{i=1}^{n} X_i = t\right] = \mathrm{E}\left[\sum_{i=1}^{n} X_i|\sum_{i=1}^{n} X_i = t\right] = t.$$

Puisque $X_1,\ldots,X_n$ sont i.i.d., chaque élément de la somme doit être équivalent et donc égal à t/n. On trouve que

$$\lambda^* = \frac{T}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.$$

3.15 On peut réécrire la fonction de masse de probabilité de la loi géométrique comme suit :

$$\begin{aligned}\Pr[X = x] &= \theta e^{\ln(1-\theta)x} \\ &= h(x)c(\theta)e^{\eta(\theta)t(x)},\end{aligned}$$

avec $h(x) = 1$, $c(\theta) = \theta$, $\eta(\theta) = \ln(1-\theta)$ et $t(x) = x$. La loi géométrique est donc membre de la famille exponentielle (voir la solution de l'exercice 3.13). Par conséquent, $T(X_1,\ldots,X_n) = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour le paramètre θ.

3.16 a) Puisque

$$\Pr[X_1 = x_1,\ldots,X_n = x_n] = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda n} \frac{1}{\prod_{i=1}^{n} x_i!}$$

se factorise en $g(\sum_{i=1}^{n} x_i,\lambda) = \lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda n}$ et $h(x_1,\ldots,x_n) = \frac{1}{\prod_{i=1}^{n} x_i!}$, le théorème de factorisation de Fisher–Neyman implique que $\sum_{i=1}^{n} X_i$ est une statistique exhaustive pour λ. On voit facilement avec le critère de Lehmann-Scheffé que cette statistique est exhaustive minimale.

Effectivement, en utilisant un autre échantillon aléatoire $y_1 \ldots y_n$ indépendant et identiquement distribué à l'échantillon $x_1 \ldots x_n$, on obtient que le rapport des vraisemblances devient

$$\begin{aligned}\frac{L(x_1 \ldots x_n;\lambda)}{L(y_1 \ldots y_n;\lambda)} &= \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-\lambda n} \frac{1}{\prod_{i=1}^{n} x_i!}}{\lambda^{\sum_{i=1}^{n} y_i} e^{-\lambda n} \frac{1}{\prod_{i=1}^{n} y_i!}} \\ &= \frac{\prod_{i=1}^{n} y_i!}{\prod_{i=1}^{n} x_i!} \lambda^{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}.\end{aligned}$$

Comme cette expression est indépendante de λ si et seulement si $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$ (pour que l'exposant sur λ soit 0 et que cette portion donne 1). Ainsi, par le critère de Lehmann-Scheffé, nous venons de démontrer que $T = \sum_{i=1}^{n} X_i$ est une statistique exhaustive minimale pour λ.

Note : On pourrait aussi dire que $\bar{X}_n$ est exhaustive minimale pour λ.

b) La règle du pouce pour trouver un MVUE est de trouver un estimateur qui est sans biais et qui est basé sur une statistique exhaustive obtenue par le théorème de factorisation de Fisher–Neyman. On a

$$\mathrm{E}[\bar{X}_n] = \lambda,$$

est donc sans biais et $\bar{X}_n = T/n$, où T est une statistique exhaustive pour λ, comme montré en a). Ainsi, $\bar{X}_n$ est un MVUE pour λ.

c) Dans ce cas,

$$\ln\{f(x;\lambda)\} = -\lambda + x\ln(\lambda) - \ln(x!)$$

et

$$\frac{\partial^2}{\partial\lambda^2} f(x;\lambda) = -\frac{x}{\lambda^2}.$$

Ainsi,

$$n\,\mathrm{E}\left[-\frac{\partial^2}{\partial\lambda^2}\ln\{f(X;\lambda)\}\right] = \frac{n}{\lambda^2}\,\mathrm{E}(X) = \frac{n}{\lambda}.$$

La borne inférieure de Cramér–Rao–Fréchet est donc

$$\frac{\lambda}{n} = \mathrm{var}(\bar{X}_n),$$

ce qui montre que $\bar{X}_n$ est un estimateur efficace pour λ.

3.17 Pour commencer, on a l'identité suivante :

$$\frac{\partial}{\partial\theta}\ln f(x;\theta) = \frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta} f(x;\theta)$$

qui peut être réécrite sous la forme

$$\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right) f(x;\theta) = \frac{\partial}{\partial\theta} f(x;\theta).$$

Ainsi, en dérivant de part et d'autre

$$\int_{-\infty}^{\infty} f(x;\theta)\,dx = 1,$$

par rapport à θ, on obtient

$$\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right) f(x;\theta)\,dx = 0.$$

En dérivant une seconde fois cette identité, on a alors

$$\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right)f(x;\theta)\,dx+\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)dx=0$$

ou, de manière équivalente,

$$\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2 f(x;\theta)\,dx=-\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right)f(x;\theta)\,dx,$$

soit

$$E\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right]=-E\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right].$$

3.18 On sait que $E[\bar{X}_n]=E[X]$ pour toute distribution et donc que $\bar{X}$ est toujours un estimateur sans biais de la moyenne. Pour une loi de Poisson, la moyenne est égale à λ et donc $\bar{X}$ est un estimateur sans biais de λ. Pour démontrer que la statistique est un estimateur à variance minimale, il faut montrer que $\bar{X}_n$ est une statistique exhaustive minimale pour λ. On a que, pour tout $x_1,\ldots,x_n,\in\mathbb{R}$,

$$\begin{aligned}f(x_1;\lambda)\times\cdots\times f(x_n;\lambda)&=\frac{\exp(-n\lambda)\lambda^{x_1+\cdots+x_n}}{\prod_{i=1}^n x_i!}\\&=\frac{\exp(-n\lambda)\lambda^{n\bar{x}_n}}{\prod_{i=1}^n x_i!}=g(\bar{x}_n;\lambda)h(x_1,\ldots,x_n).\end{aligned}$$

Selon le critère de Fisher–Neyman, $\bar{X}_n$ est une statistique exhaustive. De plus, cette statistique est minimale, puisque pour tout $x_1,\ldots,x_n,y_1,\ldots,y_n\in\mathbb{R}$, on a que le ratio

$$\begin{aligned}\frac{f(x_1;\lambda)\times\cdots\times f(x_n;\lambda)}{f(y_1;\lambda)\times\cdots\times f(y_n;\lambda)}&=\frac{\exp(-n\lambda)\lambda^{n\bar{x}_n}}{\prod_{i=1}^n x_i!}\frac{\prod_{i=1}^n y_i!}{\exp(-n\lambda)\lambda^{n\bar{y}_n}}\\&=\lambda^{n(\bar{x}_n-\bar{y}_n)}\frac{\prod_{i=1}^n y_i!}{\prod_{i=1}^n x_i!}\end{aligned}$$

ne dépend pas de λ si et seulement si $\bar{x}_n=\bar{y}_n$.
De façon alternative, on pourrait aussi démontrer que l'estimateur est sans biais à variance minimale en montrant qu'il est sans biais, puis en montrant que sa variance est égale à la borne inférieure de de Rao–Cramér. Or, d'une part, on a

$$\operatorname{var}[\bar{X}]=\frac{\operatorname{var}[X]}{n}=\frac{\lambda}{n}.$$

D'autre part,

$$\begin{aligned}\frac{\partial}{\partial\lambda}\ln f(x;\lambda)&=\frac{\partial}{\partial\lambda}(x\ln(\lambda)-\lambda-\ln(x!))\\&=\frac{x}{\lambda}-1\\&=\frac{x-\lambda}{\lambda}\end{aligned}$$

et donc

$$\begin{aligned} E\left[\left(\frac{\partial}{\partial\lambda}\ln f(X;\lambda)\right)^2\right] &= \frac{1}{\lambda^2}E\left[(X-\lambda)^2\right] \\ &= \frac{\text{var}[X]}{\lambda^2} \\ &= \frac{1}{\lambda}. \end{aligned}$$

Ainsi, la borne de Rao–Cramér est λ/n. On a donc démontré que $\bar{X}$ est un estimateur sans biais à variance minimale du paramètre λ de la loi de Poisson.

3.19 Si X a une distribution binomiale de paramètres $n \in \mathbb{N}$ et $0 \leq \theta \leq 1$, alors on peut représenter la variable aléatoire sous la forme $X = Y_1 + \cdots + Y_n$, où $Y_i \sim \text{Bernoulli}(\theta)$, $i = 1,\ldots,n$. Ainsi, $X/n = \bar{Y}$. Dès lors, on sait $\bar{Y}$ est un estimateur sans biais de $E[Y] = \theta$. Pour montrer qu'il est MVUE, soit on montre que $\bar{Y}$ est une statistique exhaustive minimale, ce qui est le cas selon le critère de Lehmann–Scheffé, soit on montre que la variance atteint la borne minimale de Cramer–Rao.
En effet, on a que $\text{var}[\bar{Y}] = \text{var}[Y]/n = \theta(1-\theta)/n$. De plus, si $f(y;\theta) = \theta^y(1-\theta)^{1-y}$, $y = 0,1$, est la densité d'une Bernoulli, alors

$$\begin{aligned} E\left[\left(\frac{\partial}{\partial\theta}\ln f(Y;\theta)\right)^2\right] &= E\left[\left(\frac{Y-\theta}{\theta(1-\theta)}\right)^2\right] \\ &= \frac{\text{var}[Y]}{[\theta(1-\theta)]^2} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

et donc la borne de Rao–Cramér est $\theta(1-\theta)/n = \text{var}[\bar{Y}]$. Par conséquent, $\bar{Y}$ est un estimateur sans biais à variance minimale du paramètre θ de la Bernoulli ou, de manière équivalente, X/n est un estimateur sans biais à variance minimale du paramètre θ de la binomiale.

3.20 a) On a

$$\begin{aligned} E[\omega\bar{X}_1 + (1-\omega)\bar{X}_2] &= \omega E\left[\bar{X}_1\right] + (1-\omega)E\left[\bar{X}_2\right] \\ &= \omega\mu + (1-\omega)\mu \\ &= \mu. \end{aligned}$$

b) En premier lieu,

$$\begin{aligned} \text{var}[\omega\bar{X}_1 - (1-\omega)\bar{X}_2] &= \omega^2\text{var}[\bar{X}_1] + (1-\omega)^2\text{var}[\bar{X}_2] \\ &= \frac{\omega^2\sigma_1^2}{n} + \frac{(1-\omega)^2\sigma_2^2}{n} \end{aligned}$$

Or, en résolvant l'équation

$$\frac{d}{d\omega}\text{var}[\omega\bar{X}_1 - (1-\omega)\bar{X}_2] = \frac{2\omega\sigma_1^2}{n} - \frac{2(1-\omega)\sigma_2^2}{n} = 0,$$

on trouve que $\omega = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$. La vérification des conditions de deuxième ordre (laissée en exercice) démontre qu'il s'agit bien d'un minimum.

c) Après quelques transformations algébriques, la variance de l'estimateur à son point minimum est

$$\frac{\sigma_1^2\sigma_2^2}{n(\sigma_1^2 + \sigma_2^2)}.$$

Lorsque $\omega = 1/2$, la variance de l'estimateur est

$$\text{var}\left[\frac{\bar{X}_1 - \bar{X}_2}{2}\right] = \frac{\sigma_1^2 + \sigma_2^2}{4n}.$$

L'efficacité relative est donc

$$\frac{(\sigma_1^2 + \sigma_2^2)^2}{4\sigma_1^2\sigma_2^2}$$

(ou l'inverse).

Chapitre 4

4.1 a) On a une une distribution de Poisson de paramètre θ, d'où $E[X] = \theta$. L'estimateur des moments est donc $\hat{\theta} = \bar{X}$.

b) La densité est celle d'une distribution bêta de paramètres θ et 1. Ainsi, $E[X] = \theta/(\theta + 1)$ et en posant

$$\frac{\theta}{\theta + 1} = \bar{X}$$

on trouve que l'estimateur des moments de θ est

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

c) On reconnaît la densité d'une distribution Gamma de paramètres 1 et θ. Ainsi, on sait que $E[X] = \theta$, d'où l'estimateur des moments est $\hat{\theta} = \bar{X}$.

d) Cette densité est celle de la loi de Laplace. On remarque d'abord que

$$-|x - \theta| = \begin{cases} -x + \theta & \text{pour } x \geq \theta \\ x - \theta & \text{pour } x \leq \theta \end{cases}.$$

Ainsi, on divise l'intégrale en deux selon les cas pour la valeur absolue :

$$E[X] = \frac{1}{2}\left(\int_{-\infty}^{\theta} xe^{x-\theta}\,dx + \int_{\theta}^{\infty} xe^{-x+\theta}\,dx\right).$$

On intègre ensuite par parties pour obtenir

$$\begin{aligned} E[X] &= \frac{1}{2}\left(\left[xe^{x-\theta}\right]_{-\infty}^{\theta} - \int_{-\infty}^{\theta} e^{x-\theta}\,dx - \left[xe^{-x+\theta}\right]_{\theta}^{\infty} + \int_{\theta}^{\infty} e^{-x+\theta}\,dx \right) \\ &= \frac{1}{2}(2\theta) \\ &= \theta. \end{aligned}$$

L'estimateur des moments de θ est donc $\hat{\theta} = \bar{X}$.

e) On a la densité d'une exponentielle de paramètre 1 translatée de θ vers la droite. Par conséquent, $E[X] = \theta + 1$, un résultat facile à vérifier en intégrant. En posant $\theta + 1 = \bar{X}$, on trouve facilement que $\hat{\theta} = \bar{X} - 1$.

4.2 a) La fonction génératrice des moments de $Z \sim \mathcal{N}(\mu,\sigma^2)$ est

$$M_Z(t) = \mathrm{E}[e^{tZ}] = \exp(\mu t + \sigma^2 t^2/2).$$

Ainsi,

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}[e^{Z}] = M_Z(1) = \exp(\mu + \sigma^2/2), \\ \mathrm{E}[Y^2] &= \mathrm{E}[e^{2Z}] = M_Z(2) = \exp(2\mu + \sigma^2 2^2/2). \end{aligned}$$

b) On doit résoudre les deux équations où $m_1 = \bar{Y}_n$ est égal à $\mathrm{E}[Y]$ et $m_2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$ est égal à $\mathrm{E}[Y^2]$:

$$\exp(\hat{\mu} + \hat{\sigma}^2/2) = \bar{Y}_n \tag{B.1}$$

$$\exp(2\hat{\mu} + 2\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} Y_i^2. \tag{B.2}$$

On note que

$$\begin{aligned} \exp(2\hat{\mu} + 2\hat{\sigma}^2) &= \left\{\exp(\hat{\mu} + \hat{\sigma}^2/2)\right\}^2 \exp(\hat{\sigma}^2) \\ &= \bar{Y}_n^2 \exp(\hat{\sigma}^2), \quad \text{avec (??)}. \end{aligned}$$

En remplaçant dans (B.2), on obtient que

$$\bar{Y}_n^2 \exp(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n} Y_i^2$$

est équivalent à

$$\hat{\sigma}^2 = \ln\left\{ \frac{\sum_{i=1}^{n} Y_i^2/n}{\bar{Y}_n^2} \right\}.$$

En utilisant (B.1), $\hat{\mu} + \hat{\sigma}^2/2 = \ln(\bar{Y}_n)$, donc

$$\begin{aligned}\hat{\mu} &= \ln(\bar{Y}_n) - \frac{1}{2}\ln\left\{\frac{\frac{1}{n}\sum_{i=1}^n Y_i^2}{\bar{Y}_n^2}\right\} \\ &= \ln(\bar{Y}_n) + \ln\left\{\frac{\bar{Y}_n}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\right\} \\ &= \ln\left\{\frac{\bar{Y}_n^2}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\right\}.\end{aligned}$$

Ainsi, les estimateurs des moments sont

$$\hat{\mu} = \ln\left\{\frac{\bar{Y}_n^2}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\right\} \quad \text{et} \quad \hat{\sigma}^2 = \ln\left\{\frac{\frac{1}{n}\sum_{i=1}^n Y_i^2}{\bar{Y}_n^2}\right\}.$$

c) Pour montrer la convergence, on note premièrement que, selon la Loi faible des grands nombres

$$\begin{aligned}\bar{Y} &\xrightarrow{P} \mathrm{E}[Y] = e^{\mu+\sigma^2/2} \\ \frac{1}{n}\sum_{i=1}^n Y_i^2 &\xrightarrow{P} \mathrm{E}[Y^2] = e^{2\mu+2\sigma^2}\end{aligned}$$

Puisque ln, le carré et la racine carrée sont toutes des fonctions continues, on trouve que, quand $n \to \infty$,

$$\begin{aligned}\hat{\mu} = \ln\left\{\frac{\bar{Y}_n^2}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\right\} &\xrightarrow{P} \ln\left\{\frac{e^{2\mu+\sigma^2}}{\sqrt{e^{2\mu+2\sigma^2}}}\right\} = \ln\left\{e^{2\mu+\sigma^2-\mu-\sigma^2}\right\} = \mu \\ \hat{\sigma}^2 = \ln\left\{\frac{\frac{1}{n}\sum_{i=1}^n Y_i^2}{\bar{Y}_n^2}\right\} &\xrightarrow{P} \ln\left\{\frac{e^{2\mu+2\sigma^2}}{e^{2\mu+\sigma^2}}\right\} = \ln\left\{e^{2\mu+2\sigma^2-2\mu-\sigma^2}\right\} = \sigma^2.\end{aligned}$$

Ainsi, les estimateurs sont convergents.

4.3 Pour obtenir l'estimateur des moments de θ, on pose

$$E[X] = \frac{1-\theta}{\theta} = \bar{X},$$

d'où

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

La moyenne de l'échantillon est $\bar{x} = 64/20 = 3{,}2$. On a donc

$$\hat{\theta} = \frac{1}{4{,}2} = 0{,}2381.$$

4.4 a) Il s'agit ici de trouver l'estimateur des moments du paramètre θ d'une distribution uniforme discrète sur $1,2,\ldots,\theta$, c'est-à-dire que $\Pr[X=x]=1/\theta$ pour $x=1,\ldots,\theta$. En posant

$$E[X]=\sum_{x=1}^{\theta} x\Pr[X=x]=\frac{1}{\theta}\sum_{x=1}^{\theta} x=\frac{\theta(\theta+1)}{2\theta}=\frac{\theta+1}{2}=\bar{X}_n,$$

on trouve facilement que l'estimateur des moments de θ est $\hat{\theta}=2\bar{X}-1$.

b) Avec $x_1=x_2=x_3=3$ et $x_4=12$, on a $\hat{\theta}=2(3+3+3+12)/4-1=9{,}5$. Or, cet estimateur est absurde puisque, en ayant roulé le résultat 12, on sait qu'il y a au moins douze faces sur le dé! En d'autres termes, 9,5 est une valeur de θ impossible. On constate que l'estimateur obtenu à l'aide de la méthode des moments n'est pas toujours un estimateur possible.

4.5 Les équations à résoudre sont

$$0{,}2=F(18{,}25)=\Phi\left(\frac{\ln(18{,}25)-\hat{\mu}}{\hat{\sigma}}\right),$$
$$0{,}8=F(35{,}8)=\Phi\left(\frac{\ln(35{,}8)-\hat{\mu}}{\hat{\sigma}}\right).$$

Les 20e et 80e quantiles de la loi normale standard sont -0,842 et 0,842, respectivement. Les équations deviennent

$$-0{,}842=\frac{2{,}904-\hat{\mu}}{\hat{\sigma}},$$
$$0{,}842=\frac{3{,}578-\hat{\mu}}{\hat{\sigma}}.$$

Diviser la première équation par la deuxième donne

$$-1=\frac{2{,}904-\hat{\mu}}{3{,}578-\hat{\mu}}.$$

La solution est $\hat{\mu}=3{,}241$ et $\hat{\sigma}=0{,}4$. La probabilité d'excéder 30 est estimée par

$$\widehat{\Pr(X>30)}=1-F(30;\hat{\mu},\hat{\sigma})=1-\Phi\left(\frac{\ln(30)-3{,}241}{0{,}4}\right)=1-\Phi(0{,}4)$$
$$=1-0{,}6554=0{,}3446.$$

4.6 Les équations à résoudre sont

$$0{,}2=F(100)=\frac{(100/\hat{\theta})^{\hat{\tau}}}{1+(100/\hat{\theta})^{\hat{\tau}}},$$
$$0{,}8=F(400)=\frac{(400/\hat{\theta})^{\hat{\tau}}}{1+(400/\hat{\theta})^{\hat{\tau}}}.$$

Avec la première équation, on obtient

$$0{,}2 = 0{,}8(100/\hat{\theta})^{\hat{\tau}} \text{ ou encore } \hat{\theta}^{\hat{\tau}} = 4(100)^{\hat{\tau}}.$$

En insérant le résultat dans la deuxième équation, on obtient

$$0{,}8 = \frac{\frac{400^{\hat{\tau}}}{\hat{\theta}^{\hat{\tau}}}}{1+\frac{400^{\hat{\tau}}}{\hat{\theta}^{\hat{\tau}}}} = \frac{\frac{(4\times100)^{\hat{\tau}}}{(4\times(100)^{\hat{\tau}})}}{1+\frac{(4\times100)^{\hat{\tau}}}{(4\times(100)^{\hat{\tau}})}} = \frac{4^{\hat{\tau}-1}}{1+4^{\hat{\tau}-1}}.$$

On résoud pour obtenir $\hat{\tau} = 2$ et $\hat{\theta} = 200$.

4.7 On a besoin de la $0{,}75(21) = 15{,}75$e plus petite observation. On obtient $0{,}25(13) + 0{,}75(14) = 13{,}75$.

4.8 Dans tous les cas, la fonction de vraisemblance est $L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$ et la fonction de log-vraisemblance est $l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta)$. L'estimateur du maximum de vraisemblance du paramètre θ est la solution de l'équation $l'(\theta) = 0$.

a) On a

$$L(\theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!},$$
$$l(\theta) = -n\theta + \sum_{i=1}^{n} x_i \ln(\theta) - \sum_{i=1}^{n} \ln(x_i!)$$

et

$$l'(\theta) = -n + \frac{\sum_{i=1}^{n} x_i}{\theta}.$$

En résolvant l'équation $l'(\theta) = 0$ pour θ, on trouve que l'estimateur du maximum de vraisemblance est $\hat{\theta} = \bar{X}$.

b) On a

$$L(\theta) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1},$$
$$l(\theta) = n\ln(\theta) + (\theta-1)\sum_{i=1}^{n} \ln(x_i)$$

et

$$l'(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln(x_i).$$

On trouve donc que

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln(X_i)} = -\frac{n}{\ln(X_1 \cdots X_n)}.$$

c) On a

$$L(\theta) = \theta^{-n} e^{-\sum_{i=1}^{n} x_i/\theta},$$
$$l(\theta) = -n\ln(\theta) - \frac{\sum_{i=1}^{n} x_i}{\theta}$$

et

$$l'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}.$$

On obtient que $\hat{\theta} = \bar{X}$.

d) On a

$$L(\theta) = \left(\frac{1}{2}\right)^n e^{-\sum_{i=1}^{n} |x_i - \theta|}$$

La présence de la valeur absolue rend cette fonction non différentiable en θ. On remarque que la fonction de vraisemblance sera maximisée lorsque l'expression $\sum_{i=1}^{n} |x_i - \theta|$ sera minimisée. On établit donc que $\hat{\theta} = \text{med}(X_1, \ldots, X_n)$, puisqu'on connaît le résultat suivant sur la médiane.

En général, si X est une variable aléatoire continue et a est une constante, on peut trouver le minimum de

$$\begin{aligned} E[|X - a|] &= \int_{-\infty}^{\infty} |x - a| f(x)\, dx \\ &= \int_{-\infty}^{a} (a - x) f(x)\, dx + \int_{a}^{\infty} (x - a) f(x)\, dx. \end{aligned}$$

Or,

$$\begin{aligned} \frac{d}{da} E[|X - a|] &= \int_{-\infty}^{a} f(x)\, dx + \int_{a}^{\infty} f(x)\, dx \\ &= F_X(a) - (1 - F_X(a)) \\ &= 2F_X(a) - 1 \end{aligned}$$

Par conséquent, le minimum est atteint au point a tel que $2F_X(a) - 1 = 0$, soit $F_X(a) = 1/2$. Par définition, cette valeur est la médiane de X.

e) On remarque que le support de la densité dépend du paramètre θ. La vraisemblance est

$$L(\theta) = e^{n\theta - \sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \mathbf{1}(x_i > \theta)$$

S'il y a une indicatrice qui est 0, alors la vraisemblance sera 0, elles doivent donc toutes être égales à 1 simultanément, ce qui est équivalent à

$$L(\theta) = e^{n\theta - \sum_{i=1}^{n} x_i} \mathbf{1}(\min(x_1, \ldots, x_n) > \theta).$$

La fonction $e^{n\theta - \sum_{i=1}^{n} x_i}$ est strictement croissante en fonction de θ, ce qui indique de choisir une valeur de θ la plus grande possible. Par contre, on a la contrainte $min(x_1, \ldots, x_n) > \theta$, c'est-à-dire que θ doit être plus inférieur ou égal à la plus petite valeur de l'échantillon. Par conséquent, $\hat{\theta} = \min(X_1, \ldots, X_n) = X_{(1)}$.

4.9 a) On a $X = Z + \mu$ où $Z \sim$ Exponentielle(λ). Alors,

$$\begin{aligned} F_X(x) &= \Pr[Z + \mu \leq x] \\ &= F_Z(x - \mu) \\ &= 1 - e^{-\lambda(x-\mu)}, \quad x > \mu \end{aligned}$$

et

$$f_X(x) = \lambda e^{-\lambda(x-\mu)}, \quad x > \mu.$$

b) On a simplement

$$\begin{aligned} E[X] &= E\big[Z + \mu\big] \\ &= E[Z] + \mu \\ &= \frac{1}{\lambda} + \mu \end{aligned}$$

et

$$\begin{aligned} \text{var}[X] &= \text{var}\big[Z + \mu\big] \\ &= \text{var}[Z] \\ &= \frac{1}{\lambda^2}. \end{aligned}$$

c) On a

$$L(\mu,\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n (x_i - \mu)} \prod_{i=1}^n \mathbf{1}(x_i \geq \mu)$$

$$l(\mu,\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^n (x_i - \mu) + \log(\mathbf{1}\{\min(x_1,\ldots,x_n) \geq \mu\})$$

et

$$\frac{\partial l(\mu,\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n (x_i - \mu),$$

d'où

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n (x_i - \hat{\mu})}.$$

On voit dans la fonction de vraisemblance que la fonction $e^{-\lambda \sum_{i=1}^n (x_i - \mu)}$ est strictement croissante en fonction de μ. Ainsi, il faut prendre la valeur de $\hat{\mu}$ la plus grande possible telle que $\log(\mathbf{1}\{\min(x_1,\ldots,x_n) \geq \hat{\mu}\}) = 0$. On a donc

$$\begin{aligned} \hat{\mu} &= x_{(1)} \\ \hat{\lambda} &= \frac{n}{\sum_{i=1}^n (x_i - x_{(1)})}. \end{aligned}$$

d) On a, par exemple, les résultats suivants pour une exponentielle translatée de paramètres $\mu = \lambda^{-1} = 1\,000$:

```
x <- rexp(100, rate = 0.001) + 1000
min(x)

## [1] 1033.883

100 / sum(x - min(x))

## [1] 0.001005155
```

Les estimations obtenues sont près des vraies valeurs des paramètres, même pour un relativement petit échantillon de taille 100.

4.10 Il est clair ici que, comme $f(x;\theta) = 1$, on ne pourra pas utiliser la technique habituelle pour calculer l'estimateur du maximum de vraisemblance. Il faut d'abord déterminer l'ensemble des valeurs de θ possibles selon l'échantillon obtenu. Comme toutes les données de l'échantillon doivent se trouver dans l'intervalle $[\theta - 1/2, \theta + 1/2]$, on a $\theta \geq X_{(n)} - 1/2$ et $\theta \leq X_{(1)} + 1/2$. De plus, puisque $X_{(n)} - X_{(1)} \leq 1$, on a que $X_{(n)} - 1/2 \leq \theta \leq X_{(1)} + 1/2$. Ainsi, toute statistique satisfaisant ces inégalités est un estimateur du maximum de vraisemblance de θ. On a donc que

$$X_{(n)} - \frac{1}{2} \leq T(X_1, \ldots, T_n) \leq X_{(1)} + \frac{1}{2}.$$

4.11 On a

$$L(\mu,\lambda) = \left(\frac{\lambda}{2\pi}\right)^{n/2} \left(\prod_{i=1}^{n} \frac{1}{x_i^3}\right)^{1/2} \exp\left\{-\frac{\lambda}{2} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\mu^2 x_i}\right\}.$$

Il est plus simple de trouver d'abord l'estimateur du maximum de vraisemblance du paramètre μ. On constate qu'il s'agit de la valeur qui minimise la somme dans l'exponentielle. Or,

$$\begin{aligned}
\frac{\partial}{\partial \mu} \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\mu^2 x_i} &= \sum_{i=1}^{n} \left[\frac{-2(x_i - \mu)}{\mu^2 x_i} + \frac{(x_i - \mu)^2}{x_i} \frac{-2}{\mu^3}\right] \\
&= \sum_{i=1}^{n} \left[\frac{-2\mu(x_i - \mu) - 2(x_i - \mu)^2}{\mu^3 x_i}\right] \\
&= -2 \sum_{i=1}^{n} \left[\frac{\mu x_i - \mu^2 + x_i^2 - 2x_i\mu + \mu^2}{\mu^3 x_i}\right] \\
&= -2 \sum_{i=1}^{n} \left[\frac{x_i^2 - x_i\mu}{\mu^3 x_i}\right] \\
&= -\sum_{i=1}^{n} \frac{2}{\mu^2} \left[\frac{x_i}{\mu} - 1\right]
\end{aligned}$$

En posant

$$\sum_{i=1}^{n} \left(\frac{x_i}{\mu} - 1\right) = 0,$$

on trouve que $\hat{\mu} = \bar{X}$. Pour trouver l'estimateur du maximum de vraisemblance de λ, on établit d'abord que

$$L(\hat{\mu}, \lambda) \propto \lambda^{n/2} e^{-\lambda H},$$

où

$$\begin{aligned} H &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{2\bar{x}^2 x_i} \\ &= \frac{1}{2} \sum_{i=1}^{n} \left(\frac{1}{x_i} - \frac{1}{\bar{x}} \right). \end{aligned}$$

On obtient donc

$$\begin{aligned} \hat{\lambda} &= \frac{n}{2H} \\ &= \frac{n}{\sum_{i=1}^{n} X_i^{-1} - \bar{X}^{-1}}. \end{aligned}$$

4.12 a) On a

$$\mathrm{E}[X_1] = \int_0^\infty x \frac{\alpha \theta^\alpha}{(x+\theta)^{\alpha+1}} \mathrm{d}x.$$

On pose $t = x + \theta$,

$$\begin{aligned} \mathrm{E}[X_1] &= \int_\theta^\infty (t-\theta) \frac{\alpha \theta^\alpha}{t^{\alpha+1}} \mathrm{d}t \\ &= \int_\theta^\infty \frac{\alpha \theta^\alpha}{t^\alpha} \mathrm{d}t - \int_\theta^\infty \frac{\alpha \theta^{\alpha+1}}{t^{\alpha+1}} \mathrm{d}t \\ &= \frac{-\alpha \theta^\alpha}{(\alpha-1) t^{\alpha-1}} \Big|_\theta^\infty + \frac{\alpha \theta^{\alpha+1}}{\alpha t^\alpha} \Big|_\theta^\infty, \quad \alpha > 1 \\ &= \frac{\alpha \theta}{\alpha - 1} - \theta = \frac{\theta}{\alpha - 1}. \end{aligned}$$

De même,

$$\mathrm{E}[X_1^2] = \int_0^\infty x^2 \frac{\alpha \theta^\alpha}{(x+\theta)^{\alpha+1}} \mathrm{d}x.$$

On pose $t = x + \theta$,

$$\mathrm{E}[X_1^2] = \int_\theta^\infty (t-\theta)^2 \frac{\alpha \theta^\alpha}{t^{\alpha+1}} \mathrm{d}t$$

En intégrant par parties,

$$\begin{aligned}
\mathrm{E}[X_1^2] &= \frac{-(t-\theta)^2\theta^\alpha}{t^\alpha}\bigg|_\theta^\infty + \int_\theta^\infty 2(t-\theta)\frac{\theta^\alpha}{t^\alpha}\mathrm{d}t \\
&= \frac{-2(t-\theta)\theta^\alpha}{(\alpha-1)t^{\alpha-1}}\bigg|_\theta^\infty + \int_\theta^\infty 2\frac{\theta^\alpha}{(\alpha-1)t^{\alpha-1}}\mathrm{d}t \\
&= \frac{-2\theta^\alpha}{(\alpha-1)(\alpha-2)t^{\alpha-2}}\bigg|_\theta^\infty, \quad \alpha > 2 \\
&= \frac{2\theta^\alpha}{(\alpha-1)(\alpha-2)\theta^{\alpha-2}} = \frac{2\theta^2}{(\alpha-1)(\alpha-2)}.
\end{aligned}$$

Ainsi

$$\mathrm{var}(X_1) = \frac{2\theta^2}{(\alpha-1)(\alpha-2)} - \frac{\theta^2}{(\alpha-1)^2} = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}.$$

b) On a $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ pour la moyenne échantillonnale et $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ pour la variance échantillonnale. Les estimateurs des moments sont tels que

$$\bar{X}_n = \frac{\hat{\theta}}{\hat{\alpha}-1} \tag{B.3}$$

$$S_n^2 = \frac{\hat{\alpha}\hat{\theta}^2}{(\hat{\alpha}-1)^2(\hat{\alpha}-2)} \tag{B.4}$$

En utilisant (B.3), on a $\hat{\theta} = \bar{X}_n(\hat{\alpha}-1)$. Aussi, on note que

$$\begin{aligned}
S_n^2 = \frac{\hat{\alpha}\hat{\theta}^2}{(\hat{\alpha}-1)^2(\hat{\alpha}-2)} &= \left(\frac{\hat{\theta}}{\hat{\alpha}-1}\right)^2 \frac{\hat{\alpha}}{\hat{\alpha}-2} \\
&= \bar{X}_n^2 \frac{\hat{\alpha}}{\hat{\alpha}-2}, \quad \text{avec (B.3)}.
\end{aligned}$$

On réarrange pour trouver

$$\hat{\alpha} = \frac{2S_n^2}{S_n^2 - \bar{X}_n^2}$$

et

$$\hat{\theta} = \bar{X}_n\left(\frac{2S_n^2}{S_n^2 - \bar{X}_n^2} - 1\right) = \bar{X}_n\frac{S_n^2 + \bar{X}_n^2}{S_n^2 - \bar{X}_n^2}.$$

c) La vraisemblance est

$$\mathcal{L}(\alpha,\theta) = \prod_{i=1}^n \frac{\alpha\theta^\alpha}{(x_i+\theta)^{\alpha+1}} = \frac{\alpha^n\theta^{n\alpha}}{\prod_{i=1}^n (x_i+\theta)^{\alpha+1}}$$

et la log-vraisemblance est

$$\begin{aligned}
\ell(\alpha,\theta) &= n\ln\alpha + n\alpha\ln\theta - \ln\left\{\prod_{i=1}^n (x_i+\theta)^{\alpha+1}\right\} \\
&= n\ln\alpha + n\alpha\ln\theta - (\alpha+1)\sum_{i=1}^n \ln(x_i+\theta).
\end{aligned}$$

Les dérivées partielles sont

$$\frac{\partial \ell(\alpha,\theta)}{\partial \alpha} = \frac{n}{\alpha} + n\ln\theta - \sum_{i=1}^{n}\ln(x_i+\theta)$$
$$\frac{\partial \ell(\alpha,\theta)}{\partial \theta} = \frac{n\alpha}{\theta} - (\alpha+1)\sum_{i=1}^{n}\frac{1}{x_i+\theta}.$$

Les estimateurs du maximum de vraisemblance sont tels que

$$\frac{n}{\hat{\alpha}} + n\ln\hat{\theta} - \sum_{i=1}^{n}\ln(x_i+\hat{\theta}) = 0$$
$$\frac{n\hat{\alpha}}{\hat{\theta}} - (\hat{\alpha}+1)\sum_{i=1}^{n}\frac{1}{x_i+\hat{\theta}} = 0.$$

4.13 La fonction de vraisemblance est

$$L(a,b) = \left(\frac{1}{b-a}\right)^n \mathbf{1}\{a < x_1,\ldots,x_n < b\} = \left(\frac{1}{b-a}\right)^n \mathbf{1}\{a < x_{(1)}\}\mathbf{1}\{b > x_{(n)}\}.$$

Pour maximiser cette fonction, il faut minimiser la quantité $b-a$ en choisissant une valeur de b la plus petite possible et une valeur de a la plus grande possible. Étant donné le support de la distribution, on choisit donc $\hat{a} = \min(X_1,\ldots,X_n)$ et $\hat{b} = \max(X_1,\ldots,X_n)$.

4.14 a) La fonction de vraisemblance de θ pour $x_1,\ldots,x_n$ est

$$\begin{aligned} L(\theta) &= \left(\frac{1}{2\theta+1}\right)^n \mathbf{1}(x_1 \le 2\theta+1)\times\cdots\times\mathbf{1}(x_n \le 2\theta+1) \\ &= \left(\frac{1}{2\theta+1}\right)^n \mathbf{1}\{\max(x_1,\ldots,x_n) \le 2\theta+1\} \\ &= \left(\frac{1}{2\theta+1}\right)^n \mathbf{1}\left\{\theta \ge \frac{\max(x_1,\ldots,x_n)-1}{2}\right\}. \end{aligned}$$

Puisque $\{1/(2\theta+1)\}^n$ est décroissante en θ, $L(\theta)$ est maximisée à

$$\hat{\theta}_n = \frac{\max(x_1,\ldots,x_n)-1}{2}$$

L'EMV de θ est donc

$$\hat{\theta}_n = \frac{\max(X_1,\ldots,X_n)-1}{2}.$$

b) La variance peut être calculée comme suit :

$$\mathrm{E}(X) = \int_0^{2\theta+1} x\,\frac{1}{2\theta+1}\,\mathrm{d}x = \frac{2\theta+1}{2}$$

et

$$\mathrm{E}(X^2) = \int_0^{2\theta+1} x^2\,\frac{1}{2\theta+1}\,\mathrm{d}x = \frac{(2\theta+1)^2}{3}.$$

Par conséquent,

$$\text{var}(X) = \text{E}(X^2) - \{\text{E}(X)\}^2 = \frac{(2\theta+1)^2}{3} - \frac{(2\theta+1)^2}{4} = \frac{(2\theta+1)^2}{12}.$$

Puisque l'application $g : (0,\infty) \to (0,\infty)$ donnée par

$$g(x) = \frac{(2x+1)^2}{12}$$

est strictement croissante (et donc un-pour-un), la propriété d'invariance de l'EMV donne que l'EMV de var(X) est

$$\frac{(2\hat{\theta}_n+1)^2}{12} = \frac{\{\max(X_1,\ldots,X_n)\}^2}{12}.$$

4.15 a) La distribution de X est une Bernoulli avec une restriction sur la valeur du paramètre θ. On a donc que $E[X] = \theta$. L'estimateur des moments de θ est donc $\tilde{\theta} = \bar{X}$.

Pour l'estimateur du maximum de vraisemblance, on a, en posant $y = \sum_{i=1}^n x_i$,

$$\begin{aligned} L(\theta) &= \theta^y(1-\theta)^{n-y}, \\ l(\theta) &= y\ln(\theta) + (n-y)\ln(1-\theta) \end{aligned}$$

et

$$\begin{aligned} l'(\theta) &= \frac{y}{\theta} - \frac{(n-y)}{1-\theta} \\ &= \frac{y-n\theta}{\theta(1-\theta)} \\ &= \frac{n\bar{x}-n\theta}{\theta(1-\theta)}. \end{aligned}$$

Ainsi, la log-vraisemblance est croissante pour $\theta \leq \bar{x}$ et décroissante pour $\theta > \bar{x}$ (voir la figure B.3). Le maximum est donc atteint en $\bar{x}$. Cependant, puisque $0 \leq \theta \leq 1/2$ on doit avoir $\hat{\theta} \leq 1/2$. On a donc $\hat{\theta} = \min(\bar{X}, 1/2)$.

b) Premièrement, on remarque que $Y = \sum_{i=1}^n X_i = n\bar{X} \sim \text{Binomiale}(n,\theta)$ avec $0 \leq \theta \leq 1/2$. Deuxièmement, on sait que $\text{EQM}(\hat{\theta}) = \text{var}[\hat{\theta}] + b(\hat{\theta})^2$, où $\hat{\theta}$ est un estimateur quelconque d'un paramètre θ et $b(\hat{\theta}) = E[\hat{\theta}] - \theta$ est le biais de l'estimateur.

Pour l'estimateur des moments $\tilde{\theta} = Y/n$, on a

$$\begin{aligned} \text{EQM}(\tilde{\theta}) &= \frac{\text{var}[Y]}{n^2} + b(Y/n)^2 \\ &= \frac{\theta(1-\theta)}{n}, \end{aligned}$$

puique $E[Y/n] = \theta$.

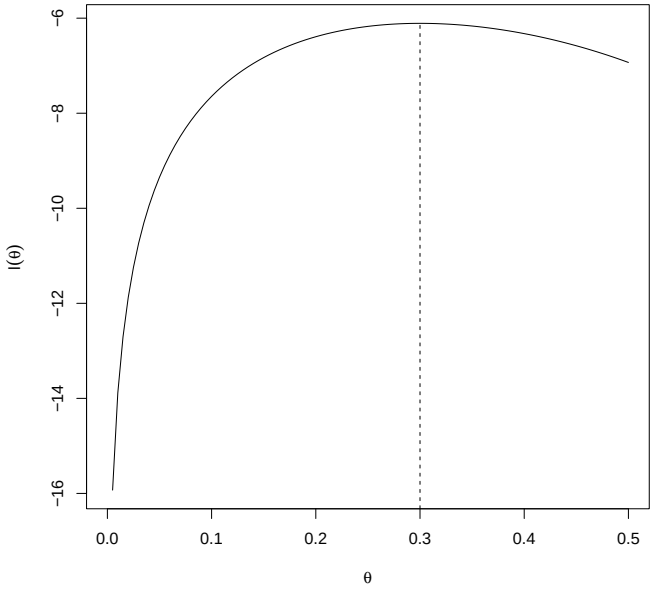


FIG. B.3 – Graphique de la fonction de log-vraisemblance de l'exercice 4.**??** pour $n = 10$ et $y = 3$.

Pour l'estimateur du maximum de vraisemblance

$$\hat{\theta} = \begin{cases} \frac{Y}{n}, & Y \leq \frac{n}{2} \\ \frac{1}{2}, & Y > \frac{n}{2}, \end{cases}$$

il est plus simple de développer l'erreur quadratique moyenne ainsi :

$$\begin{aligned} \text{EQM}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= \sum_{y=0}^{n} (\hat{\theta} - \theta)^2 \Pr[Y = y] \\ &= \sum_{y=0}^{[n/2]} \left(\frac{y}{n} - \theta\right)^2 \binom{n}{y} \theta^y (1-\theta)^{n-y} \\ &\quad + \sum_{y=[n/2]+1}^{n} \left(\frac{1}{2} - \theta\right)^2 \binom{n}{y} \theta^y (1-\theta)^{n-y}. \end{aligned}$$

c) On compare les erreurs quadratiques moyennes des deux estimateurs. Soit

$$\text{EQM}(\tilde{\theta}) = \sum_{y=0}^{n} \left(\frac{y}{n} - \theta\right)^2 \binom{n}{y} \theta^y (1-\theta)^{n-y},$$

d'où

$$\mathrm{EQM}(\tilde{\theta}) - \mathrm{EQM}(\hat{\theta}) = \sum_{y=[n/2]+1}^{n} \left(\frac{y}{n} + \frac{1}{2} - 2\theta\right)\left(\frac{y}{n} - \frac{1}{2}\right) \times \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

Étant donné que $y/n > 1/2$ et que $\theta \leq 1/2$, tous les termes dans la somme sont positifs. On a donc que $\mathrm{EQM}(\tilde{\theta}) - \mathrm{EQM}(\hat{\theta}) > 0$ ou, de manière équivalente, $\mathrm{EQM}(\hat{\theta}) < \mathrm{EQM}(\tilde{\theta})$. En terme d'erreur quadratique moyenne, l'estimateur du maximum de vraisemblance est meilleur que l'estimateur des moments selon ce critère.

4.16 a) La loi normale est définie sur les nombres réels, alors que les montants sont positifs seulement. Le domaine d'une loi normale n'est donc pas le même que le domaine des montants de réclamation. La moyenne échantillonnale et la médiane ne sont pas égales, ce qui serait le cas si le modèle normal était approprié. L'histogramme des montants ne ressemblent pas à une cloche symétrique. De bonnes options seraient les distributions Gamma, log-normale ou Pareto puisqu'elles sont toutes définies sur les réels positifs et sont asymétriques.

b) On a $\bar{x}_n = 1\,853$ et $m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 10\,438\,832$. Selon l'exercice 4.2 b), on a

$$\hat{\mu} = \ln\left\{\frac{\bar{x}_n^2}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}}\right\} = \ln\left\{\frac{1\,853^2}{\sqrt{10\,438\,832}}\right\} = 6.969$$

$$\hat{\sigma}^2 = \ln\left\{\frac{\frac{1}{n}\sum_{i=1}^{n} x_i^2}{\bar{x}_n^2}\right\} = \ln\left\{\frac{10\,438\,832}{1\,853^2}\right\} = 1.112.$$

c) On a $\bar{x}_n = 1\,853$ et $m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 10\,438\,832$, donc

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}_n^2 \\ &= \frac{n}{n-1}(m_2 - \bar{x}_n^2) \\ &= \frac{6\,773}{6\,772}(10\,438\,832 - 1\,853^2) = 7\,006\,257. \end{aligned}$$

Selon l'exercice 4.12 b), on trouve

$$\hat{\alpha} = \frac{2s_n^2}{s_n^2 - \bar{x}_n^2} = 3.922$$

$$\hat{\theta} = \bar{x}_n\left(\frac{2s_n^2}{s_n^2 - \bar{x}_n^2} - 1\right) = 5414.77.$$

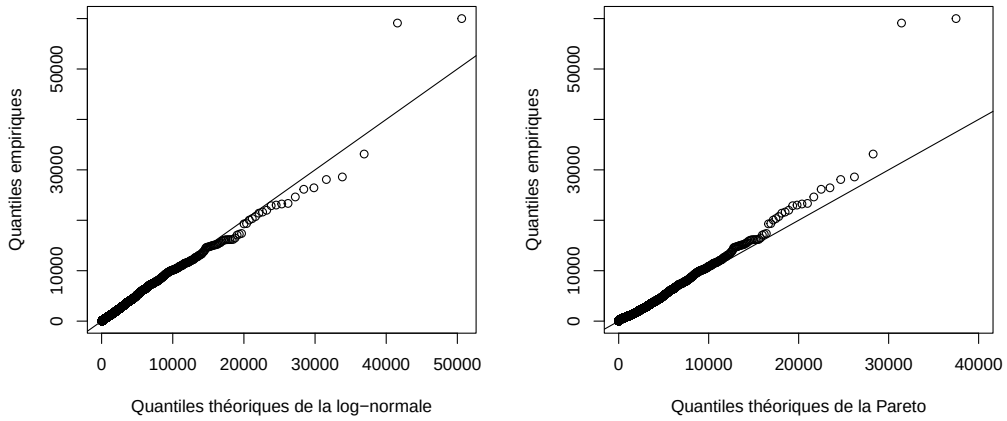
d) L'AIC est donné par $2(-\ell(\hat{\theta}) + k)$, où $\ell(\hat{\theta})$ est la valeur maximale de la vraisemblance et k le nombre de paramètres dans le modèle. Il y a deux paramètres dans chaque modèle. Les AIC sont donc

Log-normale : $2 \times (57\,185{,}11 + 2) = 114\,374{,}2$.

Pareto : $2 \times (57\,500{,}12 + 2) = 115\,004{,}2$

Selon le critère AIC, le meilleur modèle est celui avec la valeur la plus petite, la distribution log-normale est donc préférable.

e) L'ajustement des deux modèles n'est pas parfait parce que les points ne sont pas exactement alignés, plus spécialement pour de grandes valeurs de x. Les points sont au-dessus de la ligne dans le graphique Quantiles-Quantiles de la Pareto, ce qui signifie que la queue de la distribution empirique est plus épaisse que celle de la Pareto. Cela devrait inquiéter l'assureur puisque ça signifie qu'il sous-estimerait les réclamations. L'ajustement de la log-normale n'est pas parfait non plus, mais il y a moins de sous-estimation, car plusieurs points se retrouvent sous la ligne et les extrêmes sont moins éloignés. Par conséquent, on recommande l'utilisation du modèle log-normal.



f) Les quantiles de la loi Pareto sont l'inverse de sa fonction de répartition. Si ω_κ est le $(1-\kappa)$e quantile, c'est-à-dire qu'il est tel que $F(\omega_\kappa) = 1 - \kappa$, alors en utilisant la fonction de répartition fournie à l'exercice 4.12, on a

$$\begin{aligned} 1 - \left(\frac{\theta}{\omega_\kappa + \theta}\right)^\alpha &= 1 - \kappa \\ \frac{\theta}{\omega_\kappa + \theta} &= \kappa^{1/\alpha} \\ \omega_\kappa &= \frac{\theta}{\kappa^{1/\alpha}} - \theta. \end{aligned}$$

Une estimation du quantile 99 % est

$$\hat{\omega}_{1\%} = \frac{\hat{\theta}}{0{,}01^{1/\hat{\alpha}}} - \hat{\theta}$$

et en utilisant l'EMV donné en (d), on a

$$\hat{\omega}_{1\%} = \frac{6\ 819{,}891}{0{,}01^{1/4{,}71364}} - 6\ 819{,}891 = 11\ 296{,}76$$

et une estimation du quantile 99,5% est

$$\hat{\omega}_{0{,}5\%} = \frac{6\ 819{,}891}{0{,}005^{1/4{,}71364}} - 6\ 819{,}891 = 14\ 166{,}68$$

Si ν_κ représente le $(1-\kappa)$e quantile de la loi log-normale,

$$F(\nu_\kappa) = 1-\kappa \Rightarrow \Phi\left(\frac{\ln \nu_\kappa - \mu}{\sigma}\right) = 1-\kappa,$$

où Φ est la fonction de répartition de $\mathcal{N}(0,1)$. Cela signifique que le $(1-\kappa)$e quantile de la loi normale centrée réduite est utilisé pour trouver

$$\frac{\ln \nu_\kappa - \mu}{\sigma} = z_\kappa \Rightarrow \nu_\kappa = \exp(\sigma z_\kappa + \mu).$$

Une estimation du quantile 99 % est

$$\hat{\nu}_{1\%} = \exp(\hat{\sigma} z_{1\%} + \hat{\mu}),$$

où $z_{1\%} = 2{,}33$. En utilisant l'EMV donné en d), on a

$$\hat{\nu}_{1\%} = \exp(\sqrt{1{,}14698} \times 2{,}33 + 6{,}95561) = 12\ 720{,}54$$

et une estimation du quantile 99,5% est

$$\hat{\nu}_{0{,}5\%} = \exp(\sqrt{1{,}14698} \times 2{,}575 + 6{,}95561) = 16\ 537{,}1$$

Tel attendu en regardant les diagrammes quantile-quantile, les quantiles estimés par la distribution log-normale sont supérieurs à ceux estimés par la distribution Pareto.

Chapitre 5

5.1 a) La fonction génératrice des moments d'une distribution Gamma avec paramètres $\alpha = 2$ et $\beta > 0$ est donnée par

$$M_X(t) = (1-\beta t)^{-2}, \quad t < 1/\beta.$$

La fonction génératrice des moments de $T = 2(X_1 + \cdots + X_n)/\beta$ est donc

$$M_T(t) = \mathrm{E}(e^{(2t/\beta)\sum_{i=1}^n X_i}) = \prod_{i=1}^n \mathrm{E}\left\{e^{(2t/\beta)X_i}\right\} = \{M_X(2t/\beta)\}^n = (1-2t)^{-2n}$$

à condition que $2t/\beta < 1/\beta$, i.e., si $t < 1/2$. Sachant que $(1-2t)^{-2n}$ est la fonction génératrice des moments d'une distribution khi-carrée avec $4n$ degrés de liberté, $T \sim \chi^2_{(4n)}$.
Ainsi, T est une fonction de l'échantillon aléatoire et du paramètre β, avec une distribution connue qui ne dépend d'aucun paramètre inconnu : T est donc un pivot.

b) Pour construire l'intervalle de confiance bilatéral à partir de T, on note $\chi^2_{0,975,4n}$ et $\chi^2_{0,025,4n}$ les quantiles 2,5% et 97,5% d'une distribution χ^2 avec $4n$ degrés de liberté. On a donc

$$\Pr(\chi^2_{0,975,4n} \leq T \leq \chi^2_{0,025,4n}) = 0{,}95.$$

On résout les inégalités

$$\begin{aligned}
\chi^2_{0,975,4n} \leq T &\leq \chi^2_{0,025,4n} \\
\chi^2_{0,975,4n} \leq \frac{2}{\beta}\sum_{i=1}^{n} X_i &\leq \chi^2_{0,025,4n} \\
\frac{\chi^2_{0,975,4n}}{2\sum_{i=1}^{n} X_i} \leq \frac{1}{\beta} &\leq \frac{\chi^2_{0,025,4n}}{2\sum_{i=1}^{n} X_i} \\
\frac{2n\bar{X}_n}{\chi^2_{0,025,4n}} \leq \beta &\leq \frac{2n\bar{X}_n}{\chi^2_{0,975,4n}},
\end{aligned}$$

ce qui donne un l'intervalle de confiance bilatéral de niveau 95 % pour β :

$$\left[\frac{2n\bar{X}_n}{\chi^2_{0,025,4n}}, \frac{2n\bar{X}_n}{\chi^2_{0,975,4n}}\right].$$

c) Avec $n = 5$, les quantiles requis d'une distribution khi-carrée avec $4n = 20$ degrés de liberté sont

$$\chi^2_{0,975,20} = 9.59, \quad \chi^2_{0,025,20} = 34.17.$$

Sachant que $\bar{x}_n = 5{,}6$, l'intervalle de confiance de niveau 95 % pour β est

$$\left[\frac{2 \times 5 \times 5{,}6}{34.17}, \frac{2 \times 5 \times 5{,}6}{9.59}\right] = [1.64, 5.84].$$

5.2 On cherche deux statistiques L et U tel que

$$\Pr[L \leq \mu \leq U] = 1 - \alpha.$$

On sait que si $X_1, \ldots, X_n$ est un échantillon aléatoire tiré d'une distribution $N(\mu, \sigma^2)$, alors $\bar{X} \sim N(\mu, \sigma^2/n)$ ou, de manière équivalente, que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Par conséquent,

$$\Pr\left[-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\alpha/\sqrt{n}} \leq z_{\alpha/2}\right] = 1 - \alpha,$$

d'où

$$\Pr\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right] = 1 - \alpha.$$

Les statistiques L et U sont dès lors connues : $L = \bar{X} - \sigma z_{\alpha/2}/\sqrt{n}$ et $U = \bar{X} + \sigma z_{\alpha/2}/\sqrt{n}$. Un estimateur par intervalle de μ est donc

$$(\bar{X} - \sigma z_{\alpha/2}/\sqrt{n}, \bar{X} + \sigma z_{\alpha/2}/\sqrt{n}).$$

Avec $n = 20$, $\sigma^2 = 80$ et $\bar{x} = 81{,}2$, on obtient l'intervalle (77,28, 85,12).

5.3 On a $X \sim N(\mu,9)$. Tel que démontré à l'exercice 5.2,

$$\Pr\left[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{0,05} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{0,05}\right] = 0{,}90.$$

Pour satisfaire la relation $\Pr[\bar{X} - 1 < \mu < \bar{X} + 1] = 0{,}90$, on doit donc choisir

$$\frac{\sigma}{\sqrt{n}} z_{0,05} = \frac{3(1{,}645)}{\sqrt{n}} = 1.$$

On trouve que $n = 24{,}35$. On doit donc choisir une taille d'échantillon de 24 ou 25.

5.4 On sait que

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

et que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Ainsi, on peut établir que

$$\Pr\left[\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2}\right] = 1 - \alpha.$$

et qu'un intervalle de confiance de niveau $1 - \alpha$ pour μ est

$$(\bar{X} - St_{\alpha/2}/\sqrt{n},\, \bar{X} + St_{\alpha/2}/\sqrt{n}).$$

Avec $n = 17$, $\bar{x} = 4{,}7$, $s^2 = 5{,}76$ et $\alpha = 0{,}10$, on trouve que $\mu \in (3{,}7, 5{,}7)$.
Pour la variance, on cherche des valeurs a et b, $a \leq b$ tel que

$$\Pr\left[a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right] = \Pr\left[\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right] = 1 - \alpha.$$

Plusieurs valeurs de a et b satisfont cette relation. Le choix le plus simple est $a = \chi^2_{1-\alpha/2}(n-1)$ et $b = \chi^2_{\alpha/2}(n-1)$. Ainsi, un intervalle de confiance de niveau $1 - \alpha$ pour σ^2 est

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right).$$

Dans une table de la loi khi carré, on trouve que $\chi^2_{0,05}(16) = 7{,}96$ et que $\chi^2_{0,95}(16) = 26{,}30$, d'où $\sigma^2 \in (3{,}50, 11{,}58)$.

5.5 On représente la taille des étudiantes en actuariat par la variable aléatoire X et celle des étudiantes en génie civil par Y. On a

$$\bar{X} \sim N(\mu_1, \sigma_1^2/15), \qquad \bar{Y} \sim N(\mu_2, \sigma_2^2/20),$$
$$\frac{14S_1^2}{\sigma_1^2} \sim \chi^2(14), \qquad \frac{19S_2^2}{\sigma_2^2} \sim \chi^2(19)$$

et les valeurs des statistiques pour les deux échantillons sont

$$\bar{x} = 152, \qquad \bar{y} = 154,$$
$$s_1^2 = 101, \qquad s_2^2 = 112.$$

a) Si l'on suppose que $\sigma_1^2 = \sigma_2^2 = 81$, alors $\bar{X} \sim N(\mu_1, 5{,}4)$ et $\bar{Y} \sim N(\mu_2, 4{,}05)$. Par conséquent,

$$\Pr\left[-1{,}645 < \frac{\bar{X} - \mu_1}{\sqrt{5{,}4}} < 1{,}645\right] = 0{,}90$$

et

$$\Pr\left[-1{,}645 < \frac{\bar{Y} - \mu_2}{\sqrt{4{,}05}} < 1{,}645\right] = 0{,}90$$

d'où

$$152 - 1{,}645\sqrt{5{,}4} < \mu_1 < 152 + 1{,}645\sqrt{5{,}4},$$

soit $148{,}18 < \mu_1 < 155{,}82$ et

$$154 - 1{,}645\sqrt{4{,}05} < \mu_2 < 154 + 1{,}645\sqrt{4{,}05},$$

soit $150{,}69 < \mu_2 < 157{,}31$.

b) Si la variance est inconnue, on a plutôt que

$$\frac{\bar{X} - \mu_1}{S_1/\sqrt{15}} \sim t(14) \quad \text{et} \quad \frac{\bar{Y} - \mu_2}{S_2/\sqrt{20}} \sim t(19).$$

Or, $t_{0{,}05}(14) = 1{,}761$ et $t_{0{,}05}(19) = 1{,}729$, d'où

$$152 - 1{,}761\sqrt{\frac{101}{15}} < \mu_1 < 152 + 1{,}761\sqrt{\frac{101}{15}},$$

soit $147{,}43 < \mu_1 < 156{,}57$ et

$$154 - 1{,}729\sqrt{\frac{112}{20}} < \mu_2 < 154 + 1{,}729\sqrt{\frac{112}{20}},$$

soit $149{,}91 < \mu_2 < 158{,}09$.

c) On cherche un intervalle de confiance à 90 % pour la différence $\mu_1 - \mu_2$. Si 0 appartient à l'intervalle, on pourra dire que la différence entre les deux moyennes n'est pas significative à 90 %. Pour les besoins de la cause, on va supposer ici que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Or, puisque

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{15} + \frac{1}{20}}} \sim N(0,1)$$

et que

$$\frac{14S_1^2}{\sigma^2} + \frac{19S_2^2}{\sigma^2} \sim \chi^2(33)$$

on établit que

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{14S_1^2 + 19S_2^2}{33}\left(\frac{1}{15} + \frac{1}{20}\right)}} \sim t(33).$$

De plus, $t_{0,05}(33) \approx z_{0,05} = 1{,}645$, d'où l'intervalle de confiance à 90 % pour $\mu_1 - \mu_2$ est

$$\mu_1 - \mu_2 \in -2 \pm 5{,}82.$$

La différence de taille moyenne entre les deux groupes d'étudiantes n'est donc pas significative.

d) Tel que mentionné précédemment,

$$Y = \frac{14S_1^2}{\sigma_1^2} \sim \chi^2(14).$$

Or, on trouve dans une table de la loi khi carré (ou avec la fonction `qchisq` dans R) que

$$\Pr[6{,}57 < Y < 23{,}68] = 0{,}90.$$

Par conséquent,

$$\Pr\left[6{,}57 < \frac{14S_1^2}{\sigma_1^2} < 23{,}68\right] = 0{,}90$$

ou, de manière équivalente,

$$\Pr\left[\frac{14S_1^2}{23{,}68} < \sigma_1^2 < \frac{14S_1^2}{6{,}57}\right] = 0{,}90.$$

Puisque $s_1^2 = 101$ dans cet exemple, un intervalle de confiance à 90 % pour σ_1^2 est (59,71, 215,22).

e) Un peu comme en c), on détermine un intervalle de confiance pour le ratio σ_2^2/σ_1^2 et on conclut que la différence entre la variance des étudiantes en génie civil n'est pas significativement plus grande que celle des étudiantes en actuariat si cet intervalle contient la valeur 1. À la suite des conclusions en c), il est raisonnable de supposer que les moyennes des deux populations sont identiques, soit $\mu_1 = \mu_2 = \mu$. On a que

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(14,19).$$

On trouve dans une table de la loi F (ou avec la fonction `qf` dans R) que

$$\Pr[0{,}417 < F < 2{,}256] = 0{,}90.$$

Par conséquent,

$$\Pr\left[\frac{0{,}417S_2^2}{S_1^2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{2{,}256S_2^2}{S_1^2}\right] = 0{,}90$$

et un intervalle de confiance à 90 % pour σ_2^2/σ_1^2 est (0,462, 2,502). La variance σ_2^2 n'est donc pas significativement plus grande que σ_1^2.

5.6 On sait que

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Ainsi, pour des constantes a et b, $a \le b$, on a

$$\Pr\left[a \le \frac{(n-1)S^2}{\sigma^2} \le b\right] = \Pr\left[\sqrt{\frac{(n-1)S^2}{b}} \le \sigma \le \sqrt{\frac{(n-1)S^2}{a}}\right] = 1-\alpha.$$

Un estimateur par intervalle de σ est donc $(\sqrt{(n-1)S^2/b}, \sqrt{(n-1)S^2/a})$, où a et b satisfont la relation $\Pr[a \le Y \le b] = 1-\alpha$, avec $Y \sim \chi^2(n-1)$.

5.7 On sait que

$$\begin{aligned}\mu &\in \bar{X} \pm z_{0,05}\frac{\sigma}{\sqrt{n}} \\ &\in \bar{X} \pm 1{,}645\frac{5}{\sqrt{n}}.\end{aligned}$$

La longueur de l'intervalle de confiance est $2(1{,}645)(5)/\sqrt{n} = 16{,}45/\sqrt{n}$. Si l'on souhaite que $16{,}45/\sqrt{n} \le 0{,}05$, alors $n \ge 108241$.

5.8 On cherche à minimiser la longueur de l'intervalle de confiance $h(a,b) = (n-1)s^2(b-a)/(ab)$ sous la contrainte que la probabilité dans cet intervalle est $1-\alpha$, c'est-à-dire que $G(b) - G(a) = 1-\alpha$. En utilisant la méthode des multiplicateurs de Lagrange, on pose

$$L(a,b,\lambda) = \frac{(n-1)s^2}{a} - \frac{(n-1)s^2}{b} + \lambda(G(b) - G(a) - 1 + \alpha).$$

Les dérivées de cette fonction par rapport à chacune de ses variables sont :

$$\frac{\partial}{\partial a} L(a,b,\lambda) = -\frac{(n-1)s^2}{a^2} + \lambda g(a)$$
$$\frac{\partial}{\partial b} L(a,b,\lambda) = -\frac{-(n-1)s^2}{b^2} + \lambda g(b)$$
$$\frac{\partial}{\partial \lambda} L(a,b,\lambda) = G(b) - G(a) - 1 + \alpha.$$

En posant ces dérivées égales à zéro et en résolvant, on trouve que

$$\frac{g(a)}{g(b)} = \frac{b^2}{a^2}$$

ou, de manière équivalente, que $b^2 g(b) = a^2 g(a)$.

5.9 a) L'intervalle de confiance bilatéral pour μ a la forme

$$\left[\bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}, \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}}\right].$$

Sachant que $n = 20$ et $\alpha = 0{,}1$, le quantile approprié de la distribution de Student est

$$t_{19,0,05} = 1.729.$$

Avec $\bar{x}_n = 505$ et $S_n = 55$, l'intervalle devient

$$\left[505 - 1.729\,\frac{55}{\sqrt{20}}, 505 + 1.729\,\frac{55}{\sqrt{20}}\right] = [483.734, 526.266].$$

b) Puisque l'intervalle inclut 508, il n'y a pas de preuve à l'effet que la moyenne du test de langue a significativement changé depuis 2005, à un niveau de confiance 90 %.

c) L'intervalle a la même forme qu'en a) et le même quantile. Avec la moyenne échantillonnale et l'écart-type échantillonnal, l'intervalle peut être calculé comme suit

$$\left[495 - 1.729\,\frac{70}{\sqrt{20}}, 495 + 1.729\,\frac{70}{\sqrt{20}}\right] = [467.935, 522.065].$$

L'intervalle inclut 520, il n'y a donc pas une différence significative dans la moyenne du test de mathématiques depuis 2005, à un niveau de confiance 90 %.

d) Non, la méthode ne peut pas être utilisée car les échantillons ne sont pas indépendants, les tests ont été effectués par les 20 mêmes étudiants dans les deux cas.

e) L'intervalle de confiance a la forme

$$\left[\frac{n-1}{\chi^2_{n-1,\alpha/2}} S_n^2, \frac{n-1}{\chi^2_{n-1,1-\alpha/2}} S_n^2\right].$$

Les quantiles appropriés d'une distribution khi-carrée avec $n-1=19$ degrés de liberté et $\alpha=0{,}1$ sont

$$\chi^2_{n-1,1-\alpha/2}=10.117, \quad \chi^2_{n-1,\alpha/2}=30.144.$$

Avec la variance échantillonnale $S_n^2=70^2=4900$, on obtient

$$\left[\frac{19}{30.144}\times 4900, \frac{19}{10.117}\times 4900\right]=[3088.557,9202.321].$$

Une possibilité d'intervalle de confiance bilatéral à $100\times(1-\alpha)$ % pour σ est

$$\left[\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha/2}}}\,S_n, \sqrt{\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}}\,S_n\right]$$

car

$$\Pr\left(\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha/2}}}\,S_n\le\sigma\le\sqrt{\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}}\,S_n\right)$$
$$=\Pr\left(\frac{n-1}{\chi^2_{n-1,\alpha/2}}\,S_n^2\le\sigma^2\le\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}\,S_n^2\right)=1-\alpha.$$

L'intervalle devient

$$[\sqrt{3088.557},\sqrt{9202.321}]=[55.575,95.929].$$

5.10 a) Les échantillons doivent être indépendants, normalement distribués et avoir la même variance σ^2.

b) L'estimation combinée de la variance est donnée par

$$s^2=\frac{(n-1)s_n^2+(m-1)s_m^2}{n+m-2}=\frac{3\times 0{,}001+4\times 0{,}002}{4+5-2}=0.00157.$$

Le quantile 97,5% de la loi de Student avec $n+m-2=7$ degrés de liberté est

$$t_{0{,}025,7}=2.365.$$

L'intervalle de confiance à 95 % est donc donné par

$$\left[\bar{x}_n-\bar{y}_m-t_{0{,}025,7}s\sqrt{\frac{1}{n}+\frac{1}{m}}, \bar{x}_n-\bar{y}_m+t_{0{,}025,7}s\sqrt{\frac{1}{n}+\frac{1}{m}}\right]$$
$$=\left[0{,}22-0{,}17-t_{0{,}025,7}s\sqrt{\frac{1}{4}+\frac{1}{5}}, 0{,}22-0{,}17+t_{0{,}025,7}s\sqrt{\frac{1}{4}+\frac{1}{5}}\right]$$
$$=[-0.01288,0.11288].$$

c) Puisque 0 est inclus dans l'intervalle de confiance bilatéral calculé en b), les moyennes ne semblent pas différer à un seuil de 5 %.

5.11 a) Les proportions échantillonnales sont

$$p_n = \frac{126}{180} = 0{,}7, \quad q_m = \frac{54}{100} = 0{,}54.$$

Les tailles d'échantillons sont respectivement de $n = 180$ et $m = 100$. Puisque les deux sont assez grandes, une approximation d'un intervalle de confiance à 90 % peut être utilisée pour $p - q$.
D'abord, comme

$$\text{var}[p - q] = \text{var}[p] + \text{var}[q],$$

on estime que

$$\widehat{\text{var}[p - q]} = \text{var}[\bar{X}_n] + \text{var}[\bar{Y}_m] = \frac{p_n(1 - p_n)}{n} + \frac{q_m(1 - q_m)}{m}.$$

Ainsi, par la méthode de l'intervalle de confiance approximatif pour grands échantillons, on a que $\hat{\theta} \pm z_{0.05}\sqrt{\widehat{\text{var}[\theta]}}$ et on trouve

$$\left[p_n - q_m - z_{0,05}\sqrt{\frac{p_n(1 - p_n)}{n} + \frac{q_m(1 - q_m)}{m}}, p_n - q_m + z_{0,05}\sqrt{\frac{p_n(1 - p_n)}{n} + \frac{q_m(1 - q_m)}{m}}\right].$$

Le quantile 95 % d'une loi $\mathcal{N}(0,1)$ est

$$z_{0,05} = 1.645.$$

L'intervalle de confiance devient

$$\left[0.16 - 1.645\sqrt{\frac{0.7 \times 0.3}{180} + \frac{0.54 \times 0.46}{100}}, 0.16 + 1.645\sqrt{\frac{0.7 \times 0.3}{180} + \frac{0.54 \times 0.46}{100}}\right]$$
$$= [0.06062, 0.25938].$$

b) Puisque l'intervalle de confiance calculé en a) ne contient pas 0, la proportion d'enfants ainés semble être significativement plus grande dans la population d'étudiants gradués au niveau de confiance 90 %.

c) En supposant $n = m$, la taille d'échantillon de chaque groupe peut être calculée en résolvant l'équation

$$1.645\sqrt{\frac{p(1 - p)}{n} + \frac{q(1 - q)}{n}} = 0{,}05.$$

On trouve

$$n = \left(\frac{1.645}{0.05}\right)^2 \{p(1 - p) + q(1 - q)\}.$$

Sachant que pour tout $p \in (0,1)$, $p(1 - p) \leq 1/4$, une estimation conservatrice des tailles d'échantillon requises est

$$n = m = \left(\frac{1.645}{0.05}\right)^2 \times \left(\frac{1}{4} + \frac{1}{4}\right) \approx 541.205.$$

Ainsi, pour atteindre la précision requise, chaque groupe doit comprendre au moins 542 personnes.

5.12 a) La vraisemblance est

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

et la log-vraisemblance est

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Si on dérive par rapport à λ, on trouve

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

et $\ell''(\lambda) = -n/\lambda^2 < 0$, donc l'estimateur du maximum de vraisemblance est

$$\frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}_n}.$$

b) On a $\ln f(X;\lambda) = \ln \lambda - \lambda X$ donc l'information de Fisher est

$$I(\lambda) = \mathrm{E}\left[-\frac{\partial^2}{\partial \lambda^2} \ln f(X;\lambda)\right] = \mathrm{E}\left[-\frac{\partial^2}{\partial \lambda^2} (\ln \lambda - \lambda X)\right] = \mathrm{E}\left[\frac{1}{\lambda^2}\right] = \frac{1}{\lambda^2}.$$

c) La distribution limite de l'EMV est

$$\frac{\hat{\lambda}_n - \lambda}{\sqrt{1/\{nI(\lambda)\}}}$$

est asymptotiquement $\mathcal{N}(0,1)$. On peut estimer $nI(\lambda)$ au dénominateur par $nI(\hat{\lambda}) = \frac{n}{\hat{\lambda}^2}$, pour obtenir, quand $n \to \infty$,

$$\frac{\hat{\lambda} - \lambda}{\hat{\lambda}/\sqrt{n}}$$

est asymptotiquement $\mathcal{N}(0,1)$. D'une autre façon, on peut estimer $nI(\lambda)$ par

$$-\frac{\partial^2}{\partial \lambda^2} \ell(\lambda)\bigg|_{\lambda=\hat{\lambda}} = \frac{n}{\hat{\lambda}^2},$$

ce qui revient à la même réponse. Sachant que $z_{2,5\%} = 1.96$ et $\hat{\lambda} = 1/\bar{x} = 1/105{,}2$, un intervalle de confiance approximatif de niveau 95 % pour λ est

$$\begin{aligned}\left[\hat{\lambda} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}^2}{n}}, \hat{\lambda} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}^2}{n}}\right] &= \left[\frac{1}{\bar{x}_n} - z_{\alpha/2}\frac{1}{\sqrt{n}\bar{x}_n}, \frac{1}{\bar{x}_n} + z_{\alpha/2}\frac{1}{\sqrt{n}\bar{x}_n}\right] \\ &= \left[\frac{1}{105.2} - \frac{1.96}{10 \times 105.2}, \frac{1}{105.2} + \frac{1.96}{10 \times 105.2}\right] \\ &= [0.00764, 0.01137].\end{aligned}$$

d) Le paramètre d'intérêt est $\theta = \Pr[X > 300] = e^{-300\lambda}$. On a

$$\Pr\left[\frac{1}{\bar{X}_n} - z_{\alpha/2}\frac{1}{\sqrt{n}\bar{X}_n} \leq \lambda \leq \frac{1}{\bar{X}_n} + z_{\alpha/2}\frac{1}{\sqrt{n}\bar{X}_n}\right] \approx 0{,}95.$$

Multiplier l'inégalité par -300 change le signe d'inégalité, alors que l'exponentielle est une fonction croissante, donc

$$\Pr\left[\exp\left\{-300\left(\frac{1}{\bar{X}_n} - z_{\alpha/2}\frac{1}{\sqrt{n}\bar{X}_n}\right)\right\} \geq e^{-300\lambda} \geq \exp\left\{-300\left(\frac{1}{\bar{X}_n} + z_{\alpha/2}\frac{1}{\sqrt{n}\bar{X}_n}\right)\right\}\right] \approx 0{,}95.$$

Donc, un intervalle de confiance de niveau approximatif 95 % pour $\Pr[X > 300]$ est

$$[0.03302, 0.10099].$$

Chapitre 6

6.1 a) L'hypothèse nulle est simple étant donné que $\Theta_0 = \{0{,}2\}$ ne contient qu'une seule valeur. La contre-hypothèse est composite étant donné que $\Theta_1 = [0, 0{,}2) \cup (0{,}2, 1]$ contient plus d'une valeur de θ.

b) Les mesures $X_1, \ldots, X_{20}$ sont des variables aléatoires Bernoulli mutuellement indépendantes. La région critique est donc donnée par

$$\mathcal{C} = \{(x_1, \ldots, x_{20}) \in \{0,1\}^n : x_1 + \cdots + x_n \in \{0,1\} \cup \{7, \ldots, 20\}\}.$$

c) Parce que l'hypothèse nulle est simple, la taille du test et la probabilité d'erreur de type I sont les mêmes.

$$\alpha = \Pr(X \leq 1, X \geq 7 | p = 0{,}2).$$

Sous l'hypothèse nulle, $X = X_1 + \cdots + X_{20}$ est une distribution binomiale de taille $n = 20$ et probabilité $p = 0{,}2$. Ainsi, α peut être calculé comme suit :

$$\begin{aligned}
\Pr(X \leq 1, X \geq 7) &= 1 - \Pr(2 \leq X \geq 6) \\
&= 1 - \sum_{x=2}^{6} \Pr(X = x) \\
&= 1 - 0{,}8441322 \\
&= 0{,}1558678
\end{aligned}$$

```
pbinom(1,size=20,prob=0.2) + 1-pbinom(6,size=20,prob=0.2)

## [1] 0.1558678
```

d) La fonction de puissance à $p_1 \in [0,1]$ arbitraire est donnée par

$$\Pi(p_1) = \Pr(X \leq 1, X \geq 7 | p = p_1)$$

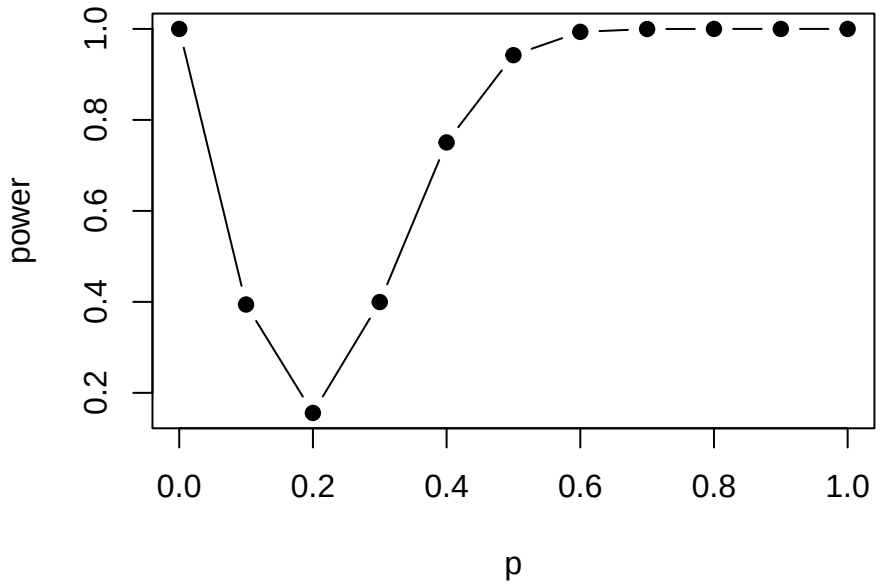
et peut être calculée par le fait que $X \sim Bin(n = 20, p = p_1)$. On répète la même démarche qu'en (c) en changeant la probabilité p de la loi Binomiale pour chacune des valeurs de p_1. Avec R, on trouve

```
p <- seq(from=0,to=1,by=0.1)
power <- pbinom(1,size=20,prob=p) + 1-pbinom(6,size=20,prob=p)
power

##  [1] 1.0000000 0.3941331 0.1558678 0.3996274 0.7505134
##  [6] 0.9423609 0.9935345 0.9997390 0.9999982 1.0000000
## [11] 1.0000000
```

La courbe de la fonction de puissance peut être tracée avec les points $(\Pi(p_1), p_1)$. En R, elle peut être tracée comme suit :

```
plot(p,power,type="b")
points(p,power,pch=16)
```



La puissance à $p = 0{,}2$ est égale à la probabilité d'erreur de type I α, alors que pour $p \neq 0{,}2$, la probabilité d'erreur de type II β est donnée par $1 - \Pi(p)$.

e) La seule différence avec (d) est que maintenant Y est une distribution binomiale de taille $n \in \{50{,}100\}$. La fonction de puissance à

$$p \in \{0, 0{,}1, 0{,}2, 0{,}3, 0{,}4, 0{,}5, 0{,}6, 0{,}7, 0{,}8, 0{,}9, 1\}$$

pour $n = 50$ et $n = 100$ est calculée de la même façon qu'en (d) en changeant la taille n de la loi Binomiale pour 50 et 100. En R, on trouve :

```
(power.50 <- pbinom(1,size=50,prob=p) + 1-pbinom(6,size=50,prob=p))
```

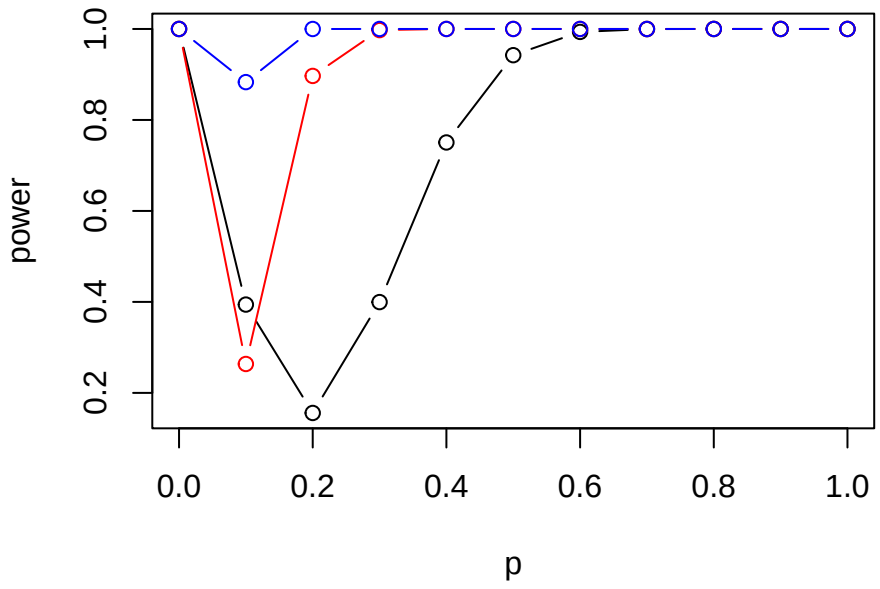
```
##  [1] 1.0000000 0.2635590 0.8967945 0.9975067 0.9999860
##  [6] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [11] 1.0000000
```

```
(power.100 <- pbinom(1,size=100,prob=p) + 1-pbinom(6,size=100,prob=p))
```

```
##  [1] 1.0000000 0.8831661 0.9999220 1.0000000 1.0000000
##  [6] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## [11] 1.0000000
```

De la même façon qu'en (d), le graphique ci-dessous trace les fonctions de puissance pour $n = 5$ (en rouge) et $n = 100$ (en bleu), avec la puissance pour $n = 20$ (en noir).

```
plot(p,power,type="b")
points(p,power.50,type="b",col="red")
points(p,power.100,type="b",col="blue")
```



Bien que la puissance se comporte bien si $p \neq 0{,}2$ (i.e., elle s'approche de 1), l'erreur de type I est complètement inacceptable : elle est aussi grande que 0,999 quand $n = 100$. Le test n'est pas du tout utile pour ces tailles d'échantillon.

f) La région de rejet devra dépendre de n, puisque plus la taille d'échantillon augmente, plus d'objets défectueux seront présents sous $\mathcal{H}_0$, simplement car plus d'objets sont inspectés. Sous l'hypothèse nulle, le nombre total d'objets défectueux va être une distribution binomiale de taille n avec probabilité $p = 0{,}2$. Les valeurs critiques pourraient être choisies comme des quantiles de la distribution binomiale.

6.2 a) On a

$$\begin{aligned} f(x_1,\ldots,x_n;\theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \theta^n \prod_{i=1}^{n} x_i^{\theta-1} \\ &= \left(\theta^n \prod_{i=1}^{n} x_i^{\theta}\right)\left(\prod_{i=1}^{n} x_i^{-1}\right) \\ &= g(t(x_1,\ldots,x_n);\theta)h(x_1,\ldots,x_n), \end{aligned}$$

où

$$\begin{aligned} g(y;\theta) &= \theta^n y^n \\ t(x_1,\ldots,x_n) &= \prod_{i=1}^{n} x_i \\ h(x_1,\ldots,x_n) &= \prod_{i=1}^{n} x_i^{-1}. \end{aligned}$$

Ainsi, par le théorème de factorisation de Fisher–Neyman, $T(X_1,\ldots,X_n) = \prod_{i=1}^{n} X_i$ est une statistique exhaustive pour θ.

b) D'une part, l'erreur de type I consiste à rejeter l'hypothèse $\mathcal{H}_0$ alors qu'elle est vraie. La probabilité de faire ce type d'erreur, notée α, correspond donc à la probabilité que la statistique du test se retrouve dans la région critique lorsque l'hypothèse $\mathcal{H}_0$ est vraie. On a donc

$$\begin{aligned} \alpha &= \Pr\left[X_1 X_2 \geq \frac{3}{4}; \theta = 1\right] \\ &= \iint_C f_{X_1X_2}(x_1,x_2;1)\,dx_2 dx_1, \end{aligned}$$

où $C = \{(x_1,x_2); x_1x_2 \geq 3/4\}$. Or,

$$f_{X_1X_2}(x_1,x_2;\theta) = f_{X_1}(x_1;\theta) f_{X_2}(x_2;\theta) = \theta^2 x_1^{\theta-1} x_2^{\theta-1},$$

d'où $f_{X_1X_2}(x_1,x_2;1) = 1$, $0 < x_1, x_2 < 1$. La région critique C est représentée à la figure B.4. Ainsi,

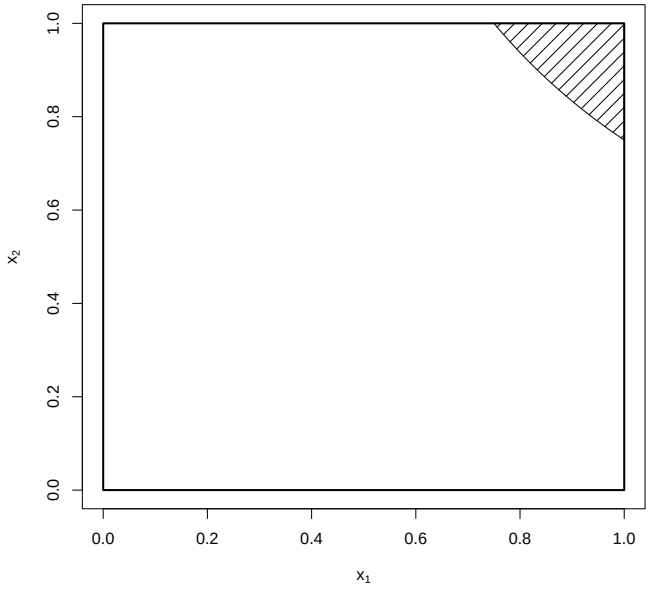


FIG. B.4 – Domaine de définition de la densité conjointe des variables X_1 et X_2 de l'exercice 6.2. La zone hachurée est la région critique $C = \{(x,y); xy \geq 3/4\}$ du test d'hypothèse.

$$\begin{aligned}
\alpha &= \int_{3/4}^{1}\int_{3/(4x_1)}^{1} dx_2\, dx_1 \\
&= \int_{3/4}^{1}\left(1 - \frac{3}{4x_1}\right) dx_1 \\
&= \frac{1}{4} + \frac{3}{4}\ln\frac{3}{4} \\
&\approx 0{,}034.
\end{aligned}$$

D'autre part, l'erreur de type II consiste à ne pas rejeter l'hypothèse $\mathcal{H}_0$ alors qu'elle est fausse. La valeur β est donc la probabilité que la statistique se retrouve à l'extérieur de la région critique lorsque l'hypothèse $\mathcal{H}_0$ est fausse, c'est-à-dire

$$\begin{aligned}
\beta &= \Pr\left[X_1X_2 < \frac{3}{4}; \theta = 2\right] \\
&= 1 - \Pr\left[X_1X_2 \geq \frac{3}{4}; \theta = 2\right] \\
&= 1 - \iint_C f_{X_1X_2}(x_1, x_2; 2)\, dx_2 dx_1.
\end{aligned}$$

Dès lors, puisque $f_{X_1X_2}(x_1, x_2; 2) = 4x_1x_2$,

$$\begin{aligned}\beta &= 1 - \int_{3/4}^{1}\int_{3/(4x_1)}^{1} 4x_1x_2\,dx_2\,dx_1 \\ &= \frac{9}{16} - \frac{9}{8}\ln\frac{3}{4} \\ &\approx 0{,}886.\end{aligned}$$

6.3 a) Ici, la fonction de distribution de l'échantillon est donnée par

$$f(\boldsymbol{x};\lambda) = e^{-\lambda n}\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\times\cdots\times x_n!}$$

et le rapport de vraisemblance est

$$\frac{f(\boldsymbol{x};\lambda_1)}{f(\boldsymbol{x};\lambda_0)} = e^{-\lambda_1 n+\lambda_0 n}\left(\frac{\lambda_1}{\lambda_0}\right)^{x_1+\cdots+x_n} = \exp\left\{-(\lambda_1-\lambda_0)n + n\bar{x}_n\ln\left(\frac{\lambda_1}{\lambda_0}\right)\right\}.$$

Parce que les hypothèses sont simples, le Lemme de Neyman–Pearson dit que le test optimal rejette l'hypothèse nulle si pour une valeur critique $k > 0$,

$$\frac{f(\boldsymbol{x};\lambda_1)}{f(\boldsymbol{x};\lambda_0)} > \frac{1}{k}$$

ce qui est équivalent à

$$\exp\left\{-(\lambda_1-\lambda_0)n + n\bar{x}_n\ln\left(\frac{\lambda_1}{\lambda_0}\right)\right\} > \frac{1}{k}.$$

Prendre le logarithme des deux côtés de l'inégalité donne

$$\bar{x}_n > \underbrace{\frac{-\ln(k)/n + (\lambda_1-\lambda_0)}{\ln(\lambda_1)-\ln(\lambda_0)}}_{=c}.$$

Sous l'hypothèse nulle, on note que $n\bar{X}_n = X_1 + \cdots + X_n$ est une distribution Poisson avec moyenne $n\lambda_0$. Ainsi, la valeur c devrait être choisie telle que

$$\Pr(n\bar{X}_n > nc|\lambda = \lambda_0) = \alpha$$

pour un seuil α, c'est-à-dire que nc est le quantile $1-\alpha$ de la distribution Poisson avec moyenne $n\lambda_0$. À noter que le seuil de signification ne sera peut-être pas exact étant donné que la distribution Poisson est discrète.

b) Ici, le test optimal rejette l'hypothèse nulle si

$$\frac{f(\boldsymbol{x};\lambda_1)}{f(\boldsymbol{x};\lambda_0)} = \exp\left\{-(\lambda_1-\lambda_0)n + n\bar{x}_n\ln\left(\frac{\lambda_1}{\lambda_0}\right)\right\} > 1,$$

c'est-à-dire lorsque

$$\bar{x}_n > \underbrace{\frac{\lambda_1-\lambda_0}{\ln(\lambda_1)-\ln(\lambda_0)}}_{=c^*}.$$

c) Si $n = 20$ et $\lambda_0 = 1/20$, $n\bar{X}_n$ est une loi de Poisson avec moyenne 1 sous l'hypothèse nulle. Le quantile de niveau $1 - 0{,}08 = 0{,}92$ de cette distribution est 2 puisque

x	$\Pr(n\bar{X}_n = x)$	$\Pr(n\bar{X}_n \leq x)$
0	0,368	0,368
1	0,368	0,736
2	0,184	0,920

Parce que

$$\begin{aligned} \Pr(n\bar{X}_n > nc | \lambda = \lambda_0) &= \alpha \\ 1 - \Pr(n\bar{X}_n \leq nc | \lambda = \lambda_0) &= \alpha \end{aligned}$$

la valeur critique c égale $2/20 = 0{,}1$ et le test rejette $\mathcal{H}_0$ si $n\bar{X}_n > 2$. La probabilité d'erreur de type I est donc

$$\Pr(n\bar{X}_n > 2 | \lambda = 1/20) = 0{,}08.$$

La probabilité d'erreur de type II est donnée par

$$\Pr(n\bar{X}_n \leq 2 | \lambda = 1/10) = 0{,}6766764$$

et peut être calculée par le fait que quand $\lambda = 1/10$, $n\bar{X}_n$ est une loi de Poisson avec moyenne 2.

d) Ici, la valeur de $c^\star$ est donnée par

$$\frac{\lambda_1 - \lambda_0}{\ln(\lambda_1) - \ln(\lambda_0)} = \frac{1/20}{\ln(1/10) - \ln(1/20)} = 0.07213,$$

donc le test rejette l'hypothèse nulle si

$$n\bar{X}_n > nc^* = 20 \times 0.07213 = 1.4427,$$

i.e. si $n\bar{X}_n > 1$. La probabilité d'erreur de type I est

$$\Pr(n\bar{X}_n > 1 | \lambda = 1/20) = 0.264.$$

La probabilité d'erreur de type II est

$$\Pr(n\bar{X}_n \leq 1 | \lambda = 1/10) = 0.406,$$

et la valeur minimale que peut atteindre $\alpha(\delta) + \beta(\delta)$ est

$$\alpha(\delta) + \beta(\delta) = 0.264 + 0.406 = 0.67.$$

6.4 On sait que $\bar{X} \sim \mathcal{N}(\mu, 5000^2/n)$. Or,

$$\begin{aligned} \alpha &= \Pr[\bar{X} \geq c; \mu = 30\,000] \\ &= 1 - \Phi\left(\frac{\sqrt{n}(c - 30\,000)}{5\,000}\right) \end{aligned}$$

d'où

$$z_\alpha = \frac{\sqrt{n}(c - 30\,000)}{5\,000}.$$

De même,

$$\begin{aligned}\beta &= \Pr[\bar{X} < c; \mu = 35\,000] \\ &= \Phi\left(\frac{\sqrt{n}(c - 35\,000)}{5\,000}\right),\end{aligned}$$

d'où

$$z_{1-\beta} = \frac{\sqrt{n}(c - 35\,000)}{5\,000}.$$

On trouve dans une table de la loi normale que $z_\alpha = z_{0,01} = 2{,}326$ et que $z_{1-\beta} = z_{0,98} = -2{,}05$. En résolvant pour n et c le système à deux équations et deux inconnues, on obtient $n = 19{,}15$ et $c = 32\,658$. Aux fins du test, on choisira donc une taille d'échantillon de $n = 19$ ou $n = 20$.

6.5 a) Il s'agit d'un simple test sur une moyenne. La statistique pour un petit échantillon est

$$T = \frac{\bar{X} - \mu_X}{\sqrt{S^2/n}} = \frac{\bar{X} - 3\,315}{\sqrt{S^2/11}} \sim t_{10}.$$

On rejette $\mathcal{H}_0$ si $t > t_{10,0,01} = 2{,}764$.

b) On commence par calculer les statistiques $\bar{X}$ et S_X^2,

$$\bar{X} = 3385{,}91 \quad \text{et} \quad S_X^2 = 113\,108{,}49.$$

On trouve que $t = 0{,}699 < t_{10,0,01} = 2{,}764$, on ne rejette donc pas $\mathcal{H}_0$ à un seuil de signification de 1 %.

c) On a un test unilatéral à gauche sur une variance pour lequel la statistique est

$$Y = \frac{(n-1)S^2}{\sigma^2} = \frac{(10)S^2}{525^2} \sim \chi^2_{10}.$$

On rejette $\mathcal{H}_0$ si $y < \chi^2_{10,0,95} = 3{,}94$.

d) Ici, $y = 4{,}104 > 3{,}94$. On ne rejette donc pas $\mathcal{H}_0$.

6.6 a) La statistique à utiliser est la même qu'à l'exercice 6.5.

b) On commence par calculer les statistiques $\bar{Y}$ et S_Y^2,

$$\bar{Y} = 3729{,}36 \quad \text{et} \quad S_Y^2 = 116\,388{,}85.$$

On trouve que $t = 4{,}028 > t_{10,0,01} = 2{,}764$, on rejette donc $\mathcal{H}_0$ à un seuil de signification de 1 %.

c) La statistique à utiliser est la même qu'à l'exercice 6.5.

d) On a $y = 4{,}223 > 3{,}94$. On ne rejette donc pas $\mathcal{H}_0$.

e) Les moyennes semblent être différentes entre les deux groupes, mais les variances égales. On commence par vérifier si le ratio des variances est près de 1. On teste

$$\mathcal{H}_0 : \sigma_X^2/\sigma_Y^2 = 1$$
$$\mathcal{H}_1 : \sigma_X^2/\sigma_Y^2 \neq 1$$

La statistique à utiliser pour ce test est

$$F = \frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim \mathcal{F}_{m-1,n-1}$$

Avec les données des deux numéros, on trouve que $f = 0{,}97$. À un seuil de signification de $\alpha = 10\%$, on trouve $\mathcal{F}_{10,10,0{,}05} = 2{,}98$ et $\mathcal{F}_{10,10,0{,}95} = 0{,}34$. On ne rejette donc pas $\mathcal{H}_0$ au seuil de 10%. Ainsi, on peut tester

$$\mathcal{H}_0 : \mu_X = \mu_Y$$
$$\mathcal{H}_1 : \mu_X \neq \mu_Y$$

en supposant les variances égales. La statistique utilisée pour ce test est

$$W = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{S_p} \sim t_{n+m-2},$$
$$\text{où } S_p^2 = \frac{S_X^2(n-1) + S_Y^2(m-1)}{n+m-2}.$$

Avec les données des deux numéros, on trouve que $w = -1{,}0139$. À un seuil de signification de $\alpha = 5\%$, on trouve $t_{20,0,025} = -2{,}086$ et $t_{20,0,975} = 2{,}086$. On ne rejette donc pas $\mathcal{H}_0$ au seuil de 5%. On n'a donc pas assez d'information pour conclure que les fillettes et les garçons nés au Québec ont un poids moyen différent les uns des autres au seuil $\alpha = 5\%$.

6.7 a) Nous avons un test unilatéral à gauche sur la différence entre deux moyennes. En supposant égales les variances des deux populations, on a $X \sim \mathcal{N}(\mu_X, \sigma^2)$ et $Y \sim \mathcal{N}(\mu_Y, \sigma^2)$. On sait que $\bar{X} \sim \mathcal{N}(\mu_X, \sigma^2/n)$ et $\bar{Y} \sim \mathcal{N}(\mu_Y, \sigma^2/m)$, d'où un estimateur de $\mu_X - \mu_Y$ sur lequel baser un test est

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_X - \mu_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right).$$

La variance σ^2 est toutefois inconnue. Un estimateur de ce paramètre est l'estimateur combiné, soit la moyenne pondérée des estimateurs de chaque échantillon :

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Or, $(n+m-2)S_p^2/\sigma^2 \sim \chi^2(n+m-2)$. Par conséquent,

$$T = \frac{\dfrac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{n^{-1}+m^{-1}}}}{\sqrt{\dfrac{(n+m-2)S_p^2}{\sigma^2}}}$$

$$= \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}} \sim t(m+n-2).$$

On rejette $\mathcal{H}_0 : \mu_X \geq \mu_Y$ en faveur de $\mathcal{H}_1 : \mu_X < \mu_y$ si $t \leq -t_{0,05}(n+m-2)$.

b) Avec les données de l'énoncé, la valeur de la statistique développée en a) est $t = -0{,}838$, alors que le 95e centile d'une loi t avec $13+16-2=27$ degrés de liberté est $t_{0,05}(27) = 1{,}703$. Puisque $|t| < 1{,}703$, on ne rejette pas $\mathcal{H}_0$.

c) On a $p = \Pr[T < -0{,}838]$, où $T \sim t(27)$. À l'aide de R, on trouve

```
pt(-0.838, df = 27)

## [1] 0.204694
```

Puisque $p = 0{,}2047 > 0{,}05$, on ne rejette pas $\mathcal{H}_0$. La conclusion est évidemment la même qu'en a).

d) On souhaite tester l'égalité de deux variances, c'est-à-dire $\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2$ versus $\mathcal{H}_1 : \sigma_X^2 \neq \sigma_Y^2$. Pour ce faire, on se base sur le fait que $(n-1)S_X^2/\sigma_X^2 \sim \chi^2(n-1)$ et que $(m-1)S_Y^2/\sigma_Y^2 \sim \chi^2(m-1)$. Ainsi, sous $\mathcal{H}_0$ (c'est-à-dire lorsque $\sigma_X^2 = \sigma_Y^2$),

$$F = \frac{(n-1)S_X^2/(n-1)}{(m-1)S_Y^2/(m-1)} \sim F(n-1, m-1).$$

On rejette $\mathcal{H}_0$ si la valeur de la statistique est supérieure au $100(1-\alpha/2)$e centile d'une loi F avec $n-1$ et $m-1$ degrés de liberté. Ici, on a $f = 0{,}8311$ et $f_{0,05}(12,15) = 2{,}48$, et pour l'autre côté, la valeur de $f_{0,95}(12,15)$ (ou de $f_{0,05}(15,12)$) n'est pas donnée dans la table, mais on peut la trouver avec R comme suit :

```
qf(0.05,12,15)

## [1] 0.3821387
```

Puisque $0.38 < 0{,}8311 < 2{,}48$, on ne rejette donc pas $\mathcal{H}_0$. L'hypothèse des variances égales est donc raisonnable.

6.8 a) Soit X la variable aléatoire du nombre de dentistes qui recommande le dentifrice parmi un groupe-test de 390 dentistes. On a donc que $X \sim \text{Binomiale}(390, \theta)$ et on teste

$$\mathcal{H}_0 : \theta = 0{,}75$$
$$\mathcal{H}_1 : \theta \neq 0{,}75.$$

Il s'agit d'un test bilatéral sur une proportion avec un grand échantillon. La statistique du test est

$$Z = \frac{\hat{\theta} - 0{,}75}{\sqrt{0{,}75(1-0{,}75)/390}}$$

et on rejette $\mathcal{H}_0$ si $|z| > z_{\alpha/2}$. Ici, on a $\alpha = 0{,}05$, $\hat{\theta} = 273/390$, d'où $z = -2{,}28$. Puisque $|z| > z_{0{,}025} = 1{,}96$, on juge, à un seuil de signification de 5 %, que la proportion de dentistes qui recommandent le dentifrice est suffisamment éloignée de la prétention du fabricant pour rejetter l'hypothèse $\mathcal{H}_0$.

b) Puisque $z_{0{,}005} = 2{,}576 > 2{,}28$, on ne rejette pas l'hypothèse $\mathcal{H}_0$ à un seuil de signification de 1 %.

c) Le seuil observé est le plus grand seuil de signification auquel on rejette $\mathcal{H}_0$. Ainsi,

$$p = 2\Pr[Z > 2{,}28] \approx 0{,}0226.$$

On rejette donc $\mathcal{H}_0$ avec un niveau de confiance maximal de 97,74 %.

6.9 a) Il s'agit d'un test bilatéral sur une proportion avec un grand échantillon. La statistique du test est

$$Z = \frac{\hat{\theta} - 0{,}20}{\sqrt{0{,}20(1-0{,}20)/n}}$$

et on rejette $\mathcal{H}_0$ si $|z| > z_{0{,}025} = 1{,}96$.

b) En calculant la valeur de la statistique pour chacune des proportions données, on constate que toutes les statistiques sont plus inférieures à 1,96. Aucun membre ne rejette donc l'hypothèse $\mathcal{H}_0$.

c) Étant donné que le seuil de signification est 5 %, si l'hypothèse $\mathcal{H}_0$ est vraie, on peut s'attendre à un taux de rejet de 5 %.

d) Si, en b), l'on n'a jamais rejeté l'hypothèse $\mathcal{H}_0$, c'est que 100 % des intervalles de confiance à 95 % pour θ contiennent la valeur 0,20.

e) La valeur de la statistique est

$$z = \frac{219/1124 - 0{,}20}{\sqrt{(0{,}20)(1-0{,}20)/1124}} = -0{,}4325$$

et $|z| < 1{,}96$. Donc, on ne rejette pas $\mathcal{H}_0$ à un seuil de signification de 5 %. La valeur p associée est :

$$\begin{aligned} p &= \Pr[|Z| > -0{,}4325] \\ &= 2\Pr[Z > 0{,}4325] \\ &= 0{,}6654, \end{aligned}$$

ce qui représente le seuil de signification minimal auquel il est possible de rejeter $\mathcal{H}_0$.

6.10 a) Il s'agit d'un test sur la différence entre deux proportions. Il faut commencer par construire la statistique. On a

$$X \sim \text{Binomiale}(n, \theta_1)$$
$$Y \sim \text{Binomiale}(m, \theta_2).$$

Pour n et m grands, on a, approximativement,

$$X \sim N(n\theta_1, n\theta_1(1-\theta_1))$$
$$Y \sim N(m\theta_2, m\theta_2(1-\theta_2),$$

et donc, toujours approximativement,

$$\hat{\theta}_1 = \frac{X}{n} \sim N\left(\theta_1, \frac{\theta_1(1-\theta_1)}{n}\right)$$
$$\hat{\theta}_2 = \frac{Y}{m} \sim N\left(\theta_2, \frac{\theta_2(1-\theta_2)}{m}\right).$$

Par conséquent,

$$\frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\theta_1(1-\theta_1)/n + \theta_2(1-\theta_2)/m}} \sim N(0,1).$$

Pour pouvoir calculer la valeur de cette statistique pour un échantillon aléatoire, on remplace θ_1 et θ_2 dans le radical par $\hat{\theta}_1 = X/n$ et $\hat{\theta}_2 = Y/m$, dans l'ordre. Un intervalle de confiance de niveau $1-\alpha$ pour $\theta_1 - \theta_2$ est donc

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{m}}.$$

De manière similaire, la statistique utilisée pour tester la différence entre les deux proportions θ_1 et θ_2 est

$$Z = \frac{(\hat{\theta}_1 - \hat{\theta}_2) - (\theta_1 - \theta_2)}{\sqrt{\hat{\theta}_1(1-\hat{\theta}_1)/n + \hat{\theta}_2(1-\hat{\theta}_2)/m}},$$

et on rejette $\mathcal{H}_0 : \theta_1 = \theta_2$ en faveur de $\mathcal{H}_1 : \theta_1 \neq \theta_2$ si $|z| > z_{\alpha/2}$.

Ici, on a $x = 351$, $y = 41$, $n = 605$ et $m = 800 - 605 = 195$. Ainsi, $\hat{\theta}_1 = 0{,}5802$, $\hat{\theta}_2 = 0{,}2103$ et $|z| = 10{,}44 > 1{,}96$. On rejette donc $\mathcal{H}_0$ à un seuil de signification de 5 %.

b) L'intervalle de confiance est

$$(\theta_1 - \theta_2) \in (0{,}3005, 0{,}4393).$$

Comme 0 n'appartient pas à cet intervalle, on rejette $\mathcal{H}_0$.

c) On cherche maintenant un intervalle de confiance pour la proportion de la population en faveur de l'introduction du taxe sur le tabac. On a une observation $x = 351 + 41 = 392$ d'une distribution Binomiale$(800, \theta)$, d'où $\hat{\theta} = 392/800 = 0{,}49$. Un intervalle de confiance à 95 % pour θ est

$$\begin{aligned}\theta &\in \hat{\theta} \pm 1{,}96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{800}} \\ &\in 0{,}49 \pm 1{,}96\sqrt{\frac{0{,}49(0{,}51)}{800}} \\ &\in (0{,}4555, 0{,}5246).\end{aligned}$$

Vérification avec R :

```
prop.test(351 + 41, 800, correct = FALSE)$conf.int
```

```
## [1] 0.4554900 0.5246056
## attr(,"conf.level")
## [1] 0.95
```

6.11 a) Premièrement, on ajuste le modèle Gamma aux données.

```
library(MASS)
m3 <- fitdistr(data,densfun="gamma",start=list(shape=0.5,rate=0.2))
```

L'estimation du paramètre α et l'estimation de son écart-type sont

```
m3$estimate[1]
```

```
##     shape
## 0.5199489
```

```
m3$sd[1]
```

```
##      shape
## 0.02679359
```

Le quantile 97,5% de la loi normale standard est

$$z_{0{,}025} = 1.96.$$

On trouve l'intervalle de confiance bilatéral approximatif pour α comme suit :

$$\begin{aligned}[\hat{\alpha} - z_{0{,}025}\hat{\sigma}_{\hat{\alpha}}, \hat{\alpha} + z_{0{,}025}\hat{\sigma}_{\hat{\alpha}}] &= [0.51995 - 1.96 \times 0.02679, 0.51995 + 1.96 \times 0.02679] \\ &= [0.46743, 0.57246].\end{aligned}$$

L'intervalle ne contient pas la valeur $\alpha = 1$. Il y a donc évidence, au seuil 5%, que $\alpha \neq 1$, c'est-à-dire que la distribution exponentielle n'est pas une bonne simplification du modèle Gamma.

b) Les hypothèses de test sont :

$$\mathcal{H}_0 : \alpha = 1, \mathcal{H}_1 : \alpha \neq 1$$

et la statistique de Wald est donnée par

$$w_n = \frac{\hat{\alpha} - 1}{\hat{\sigma}_{\hat{\alpha}}} = \frac{0.51995 - 1}{0.02679} = -17.91664.$$

Clairement, w_n est plus petite que

$$-z_{0,025} = -1.96.$$

Ainsi, $\mathcal{H}_0$ est rejetée à un niveau de confiance 5%. La conclusion est la même qu'en (a).

c) Pour calculer la statistique de test, on ajuste d'abord le modèle exponentiel aux données :

```
m2 <- fitdistr(data,densfun="exponential")
```

La log-vraisemblance des deux modèles est donnée par

```
m3$loglik

## [1] -865.6958

m2$loglik

## [1] -963.2785
```

ce qui donne la statistique du rapport de vraisemblance

$$w = 2\{\ell(\hat{\alpha},\hat{\beta}) - \ell(\hat{\beta})\} = 2(-865.6958 + 963.2785) = 195.1653.$$

Clairement, l'hypothèse que $\alpha = 1$ est rejetée parce que la valeur de la statistique du rapport de vraisemblance est beaucoup plus grande que le quantile 97,5% de la distribution khi-carrée avec 1 degré de liberté.

$$\chi^2_{1,0.025} = 5.024.$$

La conclusion est donc la même qu'en b) : le modèle exponentiel n'est pas une simplification adéquate du modèle Gamma pour l'ajustement des données.

6.12 a) L'hypothèse nulle est $\mathcal{H}_0$: les lignes et les colonnes sont indépendantes. On doit donc faire un test du χ^2. On calcule les totaux des lignes :

$$r_1 = 122 + 167 + 528 + 673 = 1490$$
$$r_2 = 203 + 118 + 178 + 212 = 711$$

On calcule les totaux des colonnes :

$$\begin{aligned} c_1 &= 122 + 203 = 325 \\ c_2 &= 167 + 118 = 285 \\ c_3 &= 528 + 178 = 706 \\ c_4 &= 673 + 212 = 885 \end{aligned}$$

On a $n = 2201$ et les nombre espérés dans les cellules sont calculés comme suit.

$$\begin{aligned} \widehat{E[n_{ij}]} &= r_i c_j / n \\ \widehat{E[n_{11}]} &= 1490 * 325/2201 = 220.0136302. \end{aligned}$$

On trouve donc les compte espérés

```
##         [,1]   [,2]   [,3]   [,4]
## [1,] 220.01 192.94 477.94 599.11
## [2,] 104.99  92.06 228.06 285.89
```

La statistique du χ^2 est

$$\begin{aligned} X^2 &= \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{\{n_{ij} - \widehat{E[n_{ij}]}\}^2}{\widehat{E[n_{ij}]}} \\ &= \frac{(122 - 220.01)^2}{220.01} + \cdots + \frac{(212 - 285.89)^2}{285.89} \\ &= 190.39 \end{aligned}$$

Le nombre de degrés de liberté est $1 \times 3 = 3$ et la valeur critique donnée dans la table est $\chi^2_{3,95\%} = 7.81473$. Puisque $190.39 > 7.81473$, on rejette l'hypothèse nulle au seuil de 5 %. La probabilité de survie dépend de la classe tarifaire.

b) On trouve

```
pchisq(190.39,3,lower.tail=FALSE)

## [1] 5.027618e-41
```

c) On peut tracer le diagramme en mosaïque avec les commandes suivantes. Le résultat se trouve dans la Figure B.5.

```
tableau <- matrix(c(122,203,167,118,528,178,673,212),nrow=2,
              dimnames=list(Survived=c("No","Yes"),
                            Class=c("1st","2nd","3rd","Crew")))
mosaicplot(t(tableau), color=TRUE, main="")
```

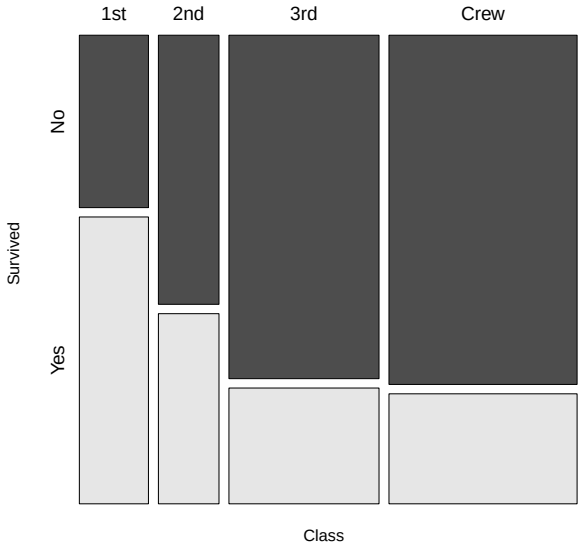


FIG. B.5 – Diagramme en mosaïque des survivants du Titanic par classe tarifaire de l'exercice 6.12.

Bibliographie

[1] J. E. Freund. *Mathematical Statistics*. Prentice Hall, Upper Saddle River, NJ, 5 edition, 1992.

[2] R.V. Hogg, A. T. Craig, and J. W. McKean. *Introduction to Mathematical Statistics*. Prentice Hall, Upper Saddle River, NJ, 6 edition, 2005.

[3] R.V. Hogg and E. A. Tanis. *Probability and Statistical Inference*. Prentice Hall, Upper Saddle River, NJ, 6 edition, 2001.

[4] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the Theory of Statistics*. McGraw-Hill, Boston, 3 edition, 1974.

[5] R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[6] D. Wackerly, W. Mendenhall, and R.L. Scheaffer. *Mathematical Statistics with Applications*. Thomson Brooks/Cole, Belmont, CA, 7 edition, 2007.

[7] D. Wackerly, W. Mendenhall, and R.L. Scheaffer. *Student's solution manual to Mathematical Statistics with Applications*. Thomson Brooks/Cole, Belmont, CA, 7 edition, 2007.