Mario
Lefebvre

Probabilités, statistique et applications

PRESSES INTERNATIONALES POLYTECHNIQUE

Mario
Lefebvre

Probabilités statistique et applications

Presses internationales
Polytechnique

Probabilités, statistique et applications
Mario Lefebvre

Couverture : Cyclone Design

Pour connaître nos distributeurs et nos points de vente, veuillez consulter notre site Web à l'adresse suivante : www.polymtl.ca/pub

Courriel des Presses internationales Polytechnique : pip@polymtl.ca

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Fonds du livre du Canada pour nos activités d'édition.

Gouvernement du Québec – Programme de crédit d'impôt pour l'édition de livres – Gestion SODEC.

Dépôt légal : 1er trimestre 2011
Bibliothèque et Archives nationales du Québec
Bibliothèque et Archives Canada

ISBN 978-2-553-01554-0 (version imprimée)
ISBN 978-2-553-01564-9 (version PDF)
Imprimé au Canada

Dieu ne joue pas aux dés.

Albert Einstein

Ils tirèrent au sort, et le sort tomba sur Matthias,
qui fut associé aux onze apôtres.

Actes 1: 26

Avant-propos

Ce livre s'adresse aux étudiants de premier cycle en sciences pures et appliquées, et particulièrement ceux en génie. Les chapitres 2 à 4 présentent la théorie des probabilités dont ces étudiants ont généralement besoin dans le cadre de leur formation. Quoique le niveau mathématique de l'exposé soit suffisamment élevé pour des non-mathématiciens, j'ai intentionnellement évité de trop entrer dans les détails. Par exemple, des sujets tels que les variables aléatoires de type mixte et la fonction delta de Dirac ne sont que brièvement mentionnés.

Les cours de probabilités sont souvent considérés comme difficiles. Cependant, après avoir enseigné cette matière pendant plusieurs années, j'en suis venu à la conclusion qu'un des principaux problèmes auxquels certains étudiants font face lorsqu'ils essaient d'apprendre la théorie des probabilités est leur faiblesse en calcul différentiel et intégral de base. Par exemple, les étudiants qui suivent un cours de probabilités ont souvent déjà oublié la technique d'intégration par parties. Pour cette raison, j'ai décidé d'inclure dans cet ouvrage un chapitre qui présente les éléments de base du calcul différentiel. Même si celui-ci ne sera probablement pas présenté en classe, les étudiants pourront s'y référer au besoin. Ce chapitre vise à donner au lecteur une bonne idée de l'utilisation en probabilités des concepts qu'il devrait déjà connaître.

Le chapitre 2 présente les résultats principaux de ce qu'on appelle les *probabilités élémentaires*, y compris la formule de Bayes et des éléments d'analyse combinatoire. Quoique ces notions ne soient pas compliquées au point de vue mathématique, c'est souvent un contenu que les étudiants ont de la difficulté à maîtriser. Il n'y a pas d'autre solution que de faire de très nombreux exercices pour se sentir à l'aise avec cette partie de la matière.

Le chapitre 3 est consacré au sujet plus technique des variables aléatoires. Tous les modèles importants pour les applications, comme les distributions binomiale et normale, y sont présentés. En général, les étudiants réussissent mieux les questions sur ce sujet dans les examens et ont l'impression que leur travail est plus récompensé que dans le cas de l'analyse combinatoire, en particulier.

Les vecteurs aléatoires, y compris le très important théorème central limite, constituent le sujet du chapitre 4. Je me suis efforcé de présenter la matière le plus simplement possible. Il reste qu'il est évident que les intégrales doubles ne peuvent pas être plus simples que les intégrales simples.

La partie *statistique* du manuel commence au chapitre 5, dans lequel les principales quantités qui permettent de caractériser un ensemble de données sont définies. Cette branche de la statistique est appelée *statistique descriptive*; elle ne devrait pas causer de problèmes aux étudiants. Le chapitre 5 traite aussi de l'estimation des paramètres des variables aléatoires, soit l'estimation ponctuelle et la technique d'estimation par intervalle de confiance.

La théorie des tests d'hypothèses est développée au chapitre 6. On présente les principaux tests d'ajustement de modèles théoriques aux données ainsi que de nombreux tests des paramètres des variables aléatoires, comme la moyenne et la variance d'une distribution gaussienne. Il s'agit d'un des principaux éléments de la statistique mathématique.

Des applications des chapitres 2 à 6 sont présentées aux chapitres 7 à 9. D'abord, le chapitre 7 traite de la régression linéaire simple, qui est un sujet de première importance en sciences appliquées. Il y est aussi question de la régression curviligne.

La fiabilité, qui intervient dans la plupart des disciplines du génie, et particulièrement en génie mécanique, fait l'objet du chapitre 8. Les notions présentées généralisent les notions de base de fiabilité vues au chapitre 2.

Enfin, les modèles de files d'attente de base sont étudiés au chapitre 9. Les étudiants en génie informatique et en génie industriel ont souvent besoin de connaissances sur ce sujet. On doit alors introduire le concept de processus stochastique, lequel est brièvement mentionné au chapitre 4 sur les vecteurs aléatoires. La théorie des files d'attente pouvant être relativement complexe, je m'en suis tenu aux modèles les plus simples. Ceux-ci sont tout de même suffisants dans la plupart des applications.

Peu importe le niveau et la formation des étudiants qui suivent un cours de probabilités et statistique, une chose est certaine: comme il est mentionné ci-dessus, il est nécessaire de résoudre plusieurs exercices avant d'avoir le sentiment d'avoir maîtrisé la théorie. À cette fin, le manuel contient près de 600 exercices, dont un grand nombre comportent plusieurs parties. À la fin de chaque chapitre, le lecteur trouvera des exercices résolus, suivis de nombreux exercices non résolus. Les réponses des exercices dont le numéro est pair sont fournies à l'appendice C. Il y a aussi plusieurs questions à choix multiple, dont les réponses sont données à l'appendice D.

Les chapitres 2 à 7 du manuel sont tirés du livre *Cours et exercices de statistique mathématique appliquée*, publié par les Presses internationales Polytechnique. Cet ouvrage est surtout axé sur la statistique, tandis que *Probabilités, statistique et applications* comporte cinq chapitres sur les probabilités (y compris les applications présentées dans les chapitres 8 et 9) et trois chapitres sur la statistique.

Il me fait plaisir de remercier toutes les personnes avec lesquelles j'ai travaillé au cours de ma carrière à l'École Polytechnique de Montréal, et qui ont fourni des exercices intéressants que j'ai inclus dans ce manuel.

Finalement, je remercie également toute l'équipe de production des Presses internationales Polytechnique.

Mario Lefebvre
Montréal, septembre 2010

Table des matières

Liste des tableaux

Liste des figures

1

Révision du calcul différentiel et intégral

Ce chapitre présente les principaux résultats du calcul différentiel et intégral utilisés en probabilités. Souvent, les étudiants qui suivent un cours sur la théorie des probabilités éprouvent de la difficulté à saisir des concepts comme les intégrales et les séries infinies. Nous rappelons en particulier la technique d'intégration par parties.

1.1 Limites et continuité

La *limite* d'une fonction est définie formellement comme suit.

Définition 1.1.1. *Soit f une fonction à valeurs réelles. On dit que $f(x)$* tend *vers f_0 ($\in \mathbb{R}$) lorsque x tend vers x_0 si, pour n'importe quel nombre positif ϵ, il existe un nombre positif δ tel que*

$$0 < |x - x_0| < \delta \quad \Longrightarrow \quad |f(x) - f_0| < \epsilon$$

On écrit: $\lim_{x \to x_0} f(x) = f_0$. *C'est-à-dire que f_0 est la* **limite** *de la fonction $f(x)$ lorsque x tend vers x_0.*

Remarques.
(i) La limite peut exister même si la fonction $f(x)$ n'est pas définie au point x_0.
(ii) Il est possible que $f(x_0)$ existe, mais que $f(x_0) \neq f_0$.
(iii) On écrit que $\lim_{x \to x_0} f(x) = \infty$ si, pour n'importe quel $M > 0$ (aussi grand que l'on veut), il existe un $\delta > 0$ tel que

$$0 < |x - x_0| < \delta \quad \Longrightarrow \quad f(x) > M$$

De façon similaire, on peut avoir $\lim_{x\to x_0} f(x) = -\infty$.

(iv) Dans la définition, on suppose que x_0 est un nombre réel. Cependant, on peut généraliser cette définition au cas où $x_0 = \pm\infty$.

Parfois, on s'intéresse à la limite de la fonction $f(x)$ lorsque x *décroît* ou *croît* vers un nombre réel donné x_0. La *limite à droite* (respectivement *limite à gauche*) de la fonction $f(x)$ lorsque x décroît (respectivement croît) vers x_0 est notée $\lim_{x\downarrow x_0} f(x)$ (respectivement $\lim_{x\uparrow x_0} f(x)$). Certains auteurs écrivent $\lim_{x\to x_0^+} f(x)$ (respectivement $\lim_{x\to x_0^-} f(x)$). Si la limite de $f(x)$ lorsque x tend vers x_0 existe, alors

$$\lim_{x\downarrow x_0} f(x) = \lim_{x\uparrow x_0} f(x) = \lim_{x\to x_0} f(x)$$

Définition 1.1.2. *On dit que la fonction à valeurs réelles $f(x)$ est* **continue** *au point $x_0 \in \mathbb{R}$ si (i) elle est définie en ce point, (ii) la limite lorsque x tend vers x_0 existe, et (iii)* $\lim_{x\to x_0} f(x) = f(x_0)$. *Si f est continue en tout point $x_0 \in [a, b]$ (ou (a, b), etc.), alors on dit que f est continue dans cet intervalle.*

Remarques.

(i) Dans ce livre, un intervalle *fermé* est noté $[a, b]$, tandis que (a, b) est un intervalle *ouvert*. On a aussi, bien sûr, les intervalles $[a, b)$ et $(a, b]$.

(ii) Si l'on écrit plutôt, dans la définition, que la limite $\lim_{x\downarrow x_0} f(x)$ (respectivement $\lim_{x\uparrow x_0} f(x)$) existe et est égale à $f(x_0)$, alors on dit que la fonction est *continue à droite* (respectivement *continue à gauche*) en x_0. Une fonction qui est continue en un certain point x_0 tel que $a < x_0 < b$ est à la fois continue à droite et continue à gauche en ce point.

(iii) On dit qu'une fonction f est *continue par morceaux* dans un intervalle $[a, b]$ si cet intervalle peut être divisé en un nombre *fini* de sous-intervalles dans lesquels f est continue et possède des limites à gauche et à droite.

(iv) Soit $f(x)$ et $g(x)$ deux fonctions à valeurs réelles. La *composition* des deux fonctions est notée $g \circ f$ et est définie par

$$(g \circ f)(x) = g[f(x)]$$

Au chapitre 3, nous utiliserons le résultat suivant: la composition de deux fonctions continues est aussi une fonction continue.

Exemple 1.1.1. Considérons la fonction

$$u(x) = \begin{cases} 0 \text{ si } x < 0 \\ 1 \text{ si } x \geq 0 \end{cases} \tag{1.1}$$

laquelle est connue sous le nom de fonction de Heaviside ou fonction échelon unitaire. En probabilités, cette fonction correspond à la *fonction de répartition* de la constante 1. Elle est aussi utilisée pour indiquer que les valeurs possibles d'une certaine *variable aléatoire* sont l'ensemble des nombres réels non négatifs. Par exemple, écrire que

$$f_X(x) = e^{-x}u(x) \quad \text{pour tout } x \in \mathbb{R}$$

est équivalent à écrire que

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

où $f_X(x)$ est appelée *fonction de densité* de la variable aléatoire X.

La fonction $u(x)$ est définie pour tout $x \in \mathbb{R}$. Dans d'autres contextes, la valeur de $u(0)$ est choisie autrement que ci-dessus. Par exemple, dans certaines applications, $u(0) = 1/2$. De toute façon, la fonction échelon unitaire est continue partout, excepté à l'origine, parce que (dans le cas présent)

$$\lim_{x \downarrow 0} u(x) = 1 \quad \text{et} \quad \lim_{x \uparrow 0} u(x) = 0$$

Cependant, avec le choix $u(0) = 1$, on peut affirmer que $u(x)$ est continue à droite en $x = 0$. ♢

Les définitions précédentes peuvent être étendues au cas des fonctions à valeurs réelles de deux (ou plusieurs) variables. En particulier, la fonction $f(x, y)$ est *continue* au point (x_0, y_0) si

$$\lim_{\substack{x \to x_0 \\ y \to y_0}} f(x, y) = f(\lim_{x \to x_0} x, \lim_{y \to y_0} y)$$

Cette formule implique que la fonction $f(x, y)$ est définie en (x_0, y_0) et que la limite de $f(x, y)$ lorsque (x, y) tend vers (x_0, y_0) existe et est égale à $f(x_0, y_0)$.

1.2 Dérivées

Définition 1.2.1. *Supposons que la fonction $f(x)$ est définie en $x_0 \in (a,b)$. Si*

$$f'(x_0) := \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \equiv \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

existe, on dit que la fonction $f(x)$ est **dérivable** *au point x_0 et que $f'(x_0)$ est la* **dérivée** *de $f(x)$ (par rapport à x) en x_0.*

Remarques.
(i) Pour que la fonction $f(x)$ soit dérivable en x_0, elle doit au moins être continue en ce point. Cependant, cette condition n'est pas suffisante, comme on peut s'en rendre compte dans l'exemple 1.2.1.

(ii) Si la limite est prise lorsque $x \downarrow x_0$ (respectivement $x \uparrow x_0$) dans la définition précédente, le résultat (si la limite existe) est appelé *dérivée à droite* (respectivement *dérivée à gauche*) de $f(x)$ en x_0, laquelle est parfois notée $f'(x_0^+)$ (respectivement $f'(x_0^-)$). Si $f'(x_0)$ existe, alors $f'(x_0^+) = f'(x_0^-)$.

(iii) La dérivée de f en un point arbitraire x est aussi notée $\frac{d}{dx} f(x)$, ou $Df(x)$. Si l'on pose $y = f(x)$, alors

$$f'(x_0) \equiv \left.\frac{dy}{dx}\right|_{x=x_0}$$

(iv) Si on dérive $f'(x)$, on obtient la *dérivée seconde* de la fonction f, notée $f''(x)$ ou $\frac{d^2}{dx^2} f(x)$. De façon similaire, $f'''(x)$ (ou $f^{(3)}(x)$, ou $\frac{d^3}{dx^3} f(x)$) est la *dérivée troisième* de f, et ainsi de suite.

(v) Une façon de trouver les valeurs de x qui maximisent ou minimisent la fonction $f(x)$ est de calculer la dérivée première $f'(x)$ et de résoudre l'équation $f'(x) = 0$. Si $f'(x_0) = 0$ et $f''(x_0) < 0$ (respectivement $f''(x_0) > 0$), alors f possède un maximum (respectivement un minimum) *relatif* en $x = x_0$. Si $f'(x) \neq 0$ pour tout $x \in \mathbb{R}$, on peut vérifier si la fonction $f(x)$ est toujours croissante ou décroissante dans l'intervalle d'intérêt.

Exemple 1.2.1. La fonction $f(x) = |x|$ est continue partout, mais n'est pas dérivable à l'origine, car on trouve que

$$f'(x_0^+) = 1 \quad \text{et} \quad f'(x_0^-) = -1$$

$\diamondsuit$

Exemple 1.2.2. La fonction $u(x)$ définie dans l'exemple 1.1.1 n'est évidemment pas dérivable en $x = 0$, parce qu'elle n'est pas continue en ce point. ♢

Exemple 1.2.3. La fonction

$$F_X(x) = \begin{cases} 0 \text{ si } x < 0 \\ x \text{ si } 0 \leq x \leq 1 \\ 1 \text{ si } x > 1 \end{cases}$$

est définie et continue partout. Elle est aussi dérivable partout, excepté en $x = 0$ et $x = 1$. On trouve que la dérivée $F'_X(x)$ de $F_X(x)$, laquelle est notée $f_X(x)$ en probabilités, est donnée par

$$f_X(x) = \begin{cases} 1 \text{ si } 0 \leq x \leq 1 \\ 0 \text{ ailleurs} \end{cases}$$

Notons que $f_X(x)$ est discontinue en $x = 0$ et $x = 1$. Plus précisément, $1 = F'(0^+) = F'(1^-)$ (tandis que $F'(0^-) = F'(1^+) = 0$). La fonction $F_X(x)$ est un exemple de *fonction de répartition* en probabilités. ♢

Remarque. En utilisant la *théorie des distributions*, on peut écrire que la dérivée de la fonction de Heaviside $u(x)$ est la *fonction delta de Dirac* $\delta(x)$ définie par

$$\delta(x) = \begin{cases} 0 \text{ si } x \neq 0 \\ \infty \text{ si } x = 0 \end{cases} \tag{1.2}$$

La fonction delta de Dirac est en fait une *fonction généralisée*. Elle est, par définition, telle que

$$\int_{-\infty}^{\infty} \delta(x)\, dx = 1$$

On a aussi, si $f(x)$ est continue en $x = x_0$, que

$$\int_{-\infty}^{\infty} f(x)\delta(x - x_0)\, dx = f(x_0)$$

Nous ne rappellerons pas les règles de dérivation de base, sauf celle qui s'applique à la *dérivée d'un quotient*:

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{g(x)f'(x) - f(x)g'(x)}{g^2(x)} \quad \text{si } g(x) \neq 0$$

Remarque. Notons que cette formule peut aussi être obtenue en dérivant le produit $f(x)h(x)$, où $h(x) := 1/g(x)$.

De même, les formules du calcul des dérivées des fonctions élémentaires sont supposées connues. Cependant, une formule qu'il vaut la peine de rappeler est la *règle de dérivation en chaîne*.

Proposition 1.2.1. (Dérivation en chaîne) *Soit $h(x)$ la fonction à valeurs réelles qui correspond à la* fonction composée $(g \circ f)(x)$. *Si f est dérivable en x et g est dérivable au point $f(x)$, alors h est aussi dérivable en x, et*

$$h'(x) = g'[f(x)]f'(x)$$

Remarque. Si l'on pose $y = f(x)$, alors la règle de dérivation en chaîne peut être écrite comme suit:

$$\frac{d}{dx}g(y) = \frac{d}{dy}g(y) \cdot \frac{dy}{dx} = g'(y)f'(x)$$

Exemple 1.2.4. Considérons la fonction $h(x) = \sqrt{x^2+1}$. On peut écrire que $h(x) = (g \circ f)(x)$, avec $f(x) = x^2 + 1$ et $g(x) = \sqrt{x}$. On a:

$$g'(x) = \frac{1}{2\sqrt{x}} \quad \Longrightarrow \quad g'[f(x)] = \frac{1}{2\sqrt{f(x)}} = \frac{1}{2\sqrt{x^2+1}}$$

Alors

$$f'(x) = 2x \quad \Longrightarrow \quad h'(x) = \frac{1}{2\sqrt{x^2+1}} \cdot (2x) = \frac{x}{\sqrt{x^2+1}}$$

♢

Finalement, un autre résultat utile est connu sous le nom de *règle de l'Hospital*.

Proposition 1.2.2. (Règle de l'Hospital) *Supposons que*

$$\lim_{x \to x_0} f(x) = \lim_{x \to x_0} g(x) = 0$$

ou que

$$\lim_{x \to x_0} f(x) = \lim_{x \to x_0} g(x) = \pm\infty$$

Si (i) $f(x)$ et $g(x)$ sont dérivables dans l'intervalle (a, b) qui contient le point x_0, sauf peut-être en x_0, et (ii) la fonction $g(x)$ est telle que $g'(x) \neq 0$ pour tout $x \neq x_0$ dans l'intervalle (a, b), alors

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = \lim_{x \to x_0} \frac{f'(x)}{g'(x)}$$

Si les fonctions $f'(x)$ et $g'(x)$ satisfont aux mêmes conditions que $f(x)$ et $g(x)$, on peut répéter le procédé. De plus, la constante x_0 peut être égale à $\pm\infty$.

Remarque. Si $x_0 = a$ ou b, on peut remplacer la limite lorsque $x \to x_0$ par $\lim x \downarrow a$ ou $\lim x \uparrow b$, respectivement.

Exemple 1.2.5. En probabilités, une façon de définir la *fonction de densité* $f_X(x)$ d'une *variable aléatoire continue* X est de calculer la limite du quotient de la *probabilité* que X prenne une valeur dans un petit intervalle autour de x par la longueur de cet intervalle:

$$f_X(x) := \lim_{\epsilon \downarrow 0} \frac{\text{Probabilité que } X \in [x - (\epsilon/2), x + (\epsilon/2)]}{\epsilon}$$

La probabilité en question est en fait égale à zéro (à la limite lorsque ϵ décroît vers 0). Par exemple, on peut avoir:

$$\text{Probabilité que } X \in [x - (\epsilon/2), x + (\epsilon/2)] = \exp\{-x + (\epsilon/2)\} - \exp\{-x - (\epsilon/2)\}$$

pour $x > 0$ et ϵ suffisamment petit.

En faisant appel à la règle de l'Hospital (avec ϵ comme variable), on trouve que

$$\begin{aligned} f_X(x) &:= \lim_{\epsilon \downarrow 0} \frac{\exp\{-x + (\epsilon/2)\} - \exp\{-x - (\epsilon/2)\}}{\epsilon} \\ &= \lim_{\epsilon \downarrow 0} \frac{\exp\{-x + (\epsilon/2)\}(1/2) - \exp\{-x - (\epsilon/2)\}(-1/2)}{1} \\ &= \exp\{-x\} \quad \text{pour } x > 0 \end{aligned}$$

♢

En deux dimensions, on définit la *dérivée partielle* de la fonction $f(x, y)$ par rapport à x par

$$\frac{\partial}{\partial x} f(x, y) \equiv f_x(x, y) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

lorsque la limite existe. La dérivée partielle de $f(x, y)$ par rapport à y est définie de façon analogue. Notons que même si les dérivées partielles $f_x(x_0, y_0)$ et $f_y(x_0, y_0)$ existent, la fonction $f(x, y)$ n'est pas nécessairement continue au point (x_0, y_0). En effet, la limite de la fonction $f(x, y)$ ne doit pas dépendre de la façon dont (x, y) tend vers (x_0, y_0). Par exemple, soit

$$f(x, y) = \begin{cases} \dfrac{xy}{x^2 + y^2} \text{ si } (x, y) \neq (0, 0) \\ \\ 0 \qquad \text{si } (x, y) = (0, 0) \end{cases} \tag{1.3}$$

Supposons que (x, y) tend vers $(0, 0)$ le long de la droite $y = kx$, où $k \neq 0$. On a alors:

$$\lim_{\substack{x \to 0 \\ y \to 0}} f(x, y) = \lim_{x \to 0} \frac{x(kx)}{x^2 + (kx)^2} = \frac{k}{1 + k^2} \quad (\neq 0)$$

Par conséquent, on doit conclure que la fonction $f(x, y)$ est discontinue en $(0, 0)$. Cependant, on peut montrer que les dérivées partielles $f_x(0, 0)$ et $f_y(0, 0)$ existent toutes les deux (et égalent 0).

1.3 Intégrales

Définition 1.3.1. *Soit $f(x)$ une fonction continue (ou continue par morceaux) dans l'intervalle $[a, b]$, et soit*

$$I = \sum_{k=1}^{n} f(\xi_k)(x_k - x_{k-1})$$

où $\xi_k \in [x_{k-1}, x_k]$ pour tout k, et $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ est une partition *de l'intervalle $[a, b]$. La limite de I lorsque n tend vers ∞, et $x_k - x_{k-1}$ décroît vers 0 pour tout k, existe et est appelée* **intégrale (définie)** *de $f(x)$ sur l'intervalle $[a, b]$. On écrit:*

$$\lim_{\substack{n \to \infty \\ (x_k - x_{k-1}) \downarrow 0 \, \forall k}} I = \int_a^b f(x)\, dx$$

Remarques.

(i) La limite ne doit pas dépendre du choix de la partition de l'intervalle $[a, b]$.

(ii) La fonction $f(x)$ est appelée *intégrande*.

(iii) Si $f(x) \geq 0$ dans l'intervalle $[a, b]$, alors l'intégrale de $f(x)$ sur $[a, b]$ donne l'aire entre la courbe $y = f(x)$ et l'axe des x de a à b.

(iv) Si l'intervalle $[a, b]$ est remplacé par un intervalle infini, ou si la fonction $f(x)$ n'est pas définie ou n'est pas bornée en au moins un point dans $[a, b]$, alors l'intégrale est dite *impropre*. Par exemple, on définit l'intégrale de $f(x)$ sur l'intervalle $[a, \infty)$ comme suit:

$$\int_a^\infty f(x)\, dx = \lim_{b \to \infty} \int_a^b f(x)\, dx$$

Si la limite existe, on dit que l'intégrale impropre est *convergente*; autrement, elle est *divergente*. Lorsque $[a, b]$ est toute la droite réelle, on devrait écrire que

$$I_1 := \int_{-\infty}^{\infty} f(x)\,dx = \lim_{a\to-\infty} \int_a^0 f(x)\,dx + \lim_{b\to\infty} \int_0^b f(x)\,dx$$

et *non pas*

$$I_1 = \lim_{b\to\infty} \int_{-b}^{b} f(x)\,dx$$

Cette dernière intégrale est en fait la *valeur principale de Cauchy* de l'intégrale I_1. La valeur principale de Cauchy peut exister même si I_1 n'existe pas.

Définition 1.3.2. *Soit $f(x)$ une fonction à valeurs réelles. Toute fonction $F(x)$ telle que $F'(x) = f(x)$ est appelée* **primitive** *(ou* **intégrale indéfinie***) de $f(x)$.*

Théorème 1.3.1. (Théorème fondamental du calcul différentiel et intégral) *Soit $f(x)$ une fonction continue dans l'intervalle $[a, b]$, et soit $F(x)$ une primitive de $f(x)$. Alors*

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

Exemple 1.3.1. La fonction

$$f_X(x) = \frac{1}{\pi(1+x^2)} \quad \text{pour } x \in \mathbb{R}$$

est la *fonction de densité* d'une *distribution de Cauchy* particulière. Pour obtenir la *valeur moyenne* de la *variable aléatoire* X, on peut calculer l'intégrale impropre

$$\int_{-\infty}^{\infty} x\,f_X(x)\,dx = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx$$

Une primitive de l'intégrande $g(x) := x\,f_X(x)$ est

$$G(x) = \frac{1}{2\pi}\ln(1+x^2)$$

On trouve que l'intégrale impropre diverge, car

$$\lim_{a\to-\infty} \int_a^0 g(x)\,dx = \lim_{a\to-\infty} [G(0) - G(a)] = -\infty$$

et

$$\lim_{b\to\infty}\int_0^b g(x)\,dx = \lim_{b\to\infty}[G(b)-G(0)] = \infty$$

Parce que $\infty-\infty$ est indéterminé, l'intégrale est effectivement divergente. Cependant, la valeur principale de Cauchy de l'intégrale est

$$\lim_{b\to\infty}\int_{-b}^b g(x)\,dx = \lim_{b\to\infty}[G(b)-G(-b)] = \lim_{b\to\infty} 0 = 0$$

♢

Il y a plusieurs résultats au sujet des intégrales que nous pourrions rappeler ici. Nous nous limitons à mentionner deux techniques qui sont utiles pour trouver des intégrales indéfinies ou pour évaluer des intégrales définies.

1.3.1 Techniques d'intégration particulières

(1) D'abord, nous rappelons au lecteur la technique connue sous le nom d'*intégration par substitution*. Soit $x = g(y)$. On peut écrire que

$$\int f(x)\,dx = \int f[g(y)]g'(y)\,dy$$

Ce résultat découle en fait de la règle de dérivation en chaîne. Dans le cas d'une intégrale définie, si l'on suppose que

(i) $f(x)$ est continue dans l'intervalle $[a,b]$,

(ii) la fonction inverse $g^{-1}(x)$ existe et

(iii) $g'(y)$ est continue dans $[g^{-1}(a), g^{-1}(b)]$ (respectivement $[g^{-1}(b), g^{-1}(a)]$) si $g(y)$ est une fonction croissante (respectivement décroissante),

on a:

$$\int_a^b f(x)\,dx = \int_c^d f[g(y)]g'(y)\,dy$$

où $a = g(c) \Leftrightarrow c = g^{-1}(a)$ et $b = g(d) \Leftrightarrow d = g^{-1}(b)$.

Exemple 1.3.2. Supposons que l'on désire évaluer l'intégrale définie

$$I_2 := \int_0^4 x^{-1/2}e^{x^{1/2}}\,dx$$

En effectuant la substitution $x = g(y) = y^2$, de sorte que $y = g^{-1}(x) = x^{1/2}$ (pour $x \in [0,4]$), on peut écrire que

$$\int_0^4 x^{-1/2} e^{x^{1/2}}\, dx \overset{y=x^{1/2}}{=} \int_0^2 y^{-1} e^y\, 2y\, dy = 2e^y \Big|_0^2 = 2(e^2 - 1)$$

♢

Remarque. Si l'on ne cherche qu'une primitive de la fonction $f(x)$, alors après avoir trouvé une primitive de $f[g(y)]g'(y)$, on doit remplacer y par $g^{-1}(x)$ (en supposant que l'inverse existe) dans la fonction obtenue. Donc, dans l'exemple ci-dessus, on a:

$$\int x^{-1/2} e^{x^{1/2}}\, dx = \int y^{-1} e^y\, 2y\, dy = 2e^y = 2e^{x^{1/2}}$$

(2) Une autre technique d'intégration très utile est basée sur la formule de la dérivée d'un produit:

$$\frac{d}{dx} f(x)g(x) = f'(x)g(x) + f(x)g'(x)$$

En intégrant les deux membres de l'équation précédente, on obtient:

$$f(x)g(x) = \int f'(x)g(x)dx + \int f(x)g'(x)dx$$

$$\iff \quad \int f(x)g'(x)\, dx = f(x)g(x) - \int f'(x)g(x)dx$$

En posant $u = f(x)$ et $v = g(x)$, on peut écrire que

$$\int u dv = uv - \int v\, du$$

Cette technique est appelée *intégration par parties*. Elle est souvent utilisée en probabilités pour calculer les *moments* d'une *variable aléatoire*.

Exemple 1.3.3. Pour obtenir la *moyenne* du carré d'une *variable aléatoire gaussienne centrée réduite*, on peut essayer d'évaluer l'intégrale

$$I_3 := \int_{-\infty}^{\infty} cx^2 e^{-x^2/2} dx$$

où c est une constante positive. Lorsqu'on applique la technique d'intégration par parties, on doit décider quelle partie de l'intégrande assigner à u. Bien sûr,

la quantité u devrait être choisie de façon que la nouvelle intégrale (indéfinie) soit plus facile à trouver. Dans le cas de I_3, on pose

$$u = cx \quad \text{et} \quad dv = xe^{-x^2/2}dx$$

Puisque

$$v = \int dv = \int xe^{-x^2/2}dx = -e^{-x^2/2}$$

il s'ensuit que

$$I_3 = cx(-e^{-x^2/2})\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} ce^{-x^2/2}dx$$

La constante c est telle que l'intégrale impropre ci-dessus est égale à 1 (voir le chapitre 3). De plus, en utilisant la règle de l'Hospital, on trouve que

$$\lim_{x\to\infty} xe^{-x^2/2} = \lim_{x\to\infty} \frac{x}{e^{x^2/2}} = \lim_{x\to\infty} \frac{1}{e^{x^2/2}\cdot x} = 0$$

De même,

$$\lim_{x\to-\infty} xe^{-x^2/2} = 0$$

De là, on peut écrire que $I_3 = 0 + 1 = 1$.

Remarques.

(i) Notons qu'aucune fonction élémentaire n'est une primitive de la fonction $e^{-x^2/2}$. Par conséquent, on n'aurait pas pu obtenir une primitive (exprimée à l'aide de fonctions élémentaires) de $x^2e^{-x^2/2}$ en procédant comme ci-dessus.

(ii) Si l'on pose plutôt $u = ce^{-x^2/2}$ (et $dv = x^2dx$), alors l'intégrale résultante est plus compliquée. En effet, on a alors:

$$I_3 = ce^{-x^2/2}\,\frac{x^3}{3}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} c\frac{x^4}{3}e^{-x^2/2}dx = \int_{-\infty}^{\infty} c\frac{x^4}{3}e^{-x^2/2}dx$$

♢

Une intégrale impropre particulière qui est importante en probabilités est définie par

$$F(\omega) = \int_{-\infty}^{\infty} e^{j\omega x}f(x)dx$$

où $j := \sqrt{-1}$ et ω est un paramètre réel, et où l'on suppose que la fonction à valeurs réelles $f(x)$ est *absolument intégrable*; c'est-à-dire que

$$\int_{-\infty}^{\infty} |f(x)|dx < \infty$$

La fonction $F(\omega)$ est appelée *transformée de Fourier* (à un facteur constant près) de la fonction $f(x)$. On peut montrer que

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-j\omega x} F(\omega) d\omega$$

On dit que $f(x)$ est la *transformée de Fourier inverse* de $F(\omega)$.

En probabilités, une *fonction de densité* $f_X(x)$ est, par définition, non négative et telle que

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

De là, la fonction $F(\omega)$ est bien définie. Dans ce contexte, elle est connue sous le nom de *fonction caractéristique* de la *variable aléatoire* X et est souvent notée $C_X(\omega)$. En la dérivant, on peut obtenir les *moments* de X (généralement plus facilement qu'en effectuant les intégrales appropriées).

Exemple 1.3.4. La fonction caractéristique d'une *variable aléatoire gaussienne centrée réduite* est

$$C_X(\omega) = e^{-\omega^2/2}$$

On peut montrer que la *moyenne* du carré de X est donnée par

$$-\frac{d^2}{d\omega^2} e^{-\omega^2/2}\Big|_{\omega=0} = -e^{-\omega^2/2}(\omega^2 - 1)\Big|_{\omega=0} = 1$$

ce qui est en accord avec le résultat obtenu dans l'exemple précédent. ♢

Finalement, pour obtenir les *moments* de la *variable aléatoire* X, on peut utiliser sa *fonction génératrice des moments*, laquelle, dans le *cas continu*, est définie (si l'intégrale existe) par

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$$

où l'on peut supposer que t est un paramètre réel. Lorsque X est non négative, la fonction génératrice des moments est en fait la *transformée de Laplace* de la fonction $f_X(x)$.

1.3.2 Intégrales doubles

Au chapitre 4 portant sur les *vecteurs aléatoires*, nous devrons évaluer des intégrales doubles pour obtenir diverses quantités d'intérêt. Une *fonction de densité conjointe* est une fonction non négative $f_{X,Y}(x,y)$ telle que

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

L'évaluation d'intégrales doubles est requise, en général, pour calculer la probabilité $P[A]$ que le *vecteur aléatoire continu* (X,Y) prenne une valeur dans un sous-ensemble donné A de $\mathbb{R}^2$. On a:

$$P[A] = \int\limits_A \int f_{X,Y}(x,y)\,dxdy$$

On doit décrire la région A à l'aide de fonctions de x ou de y (selon ce qui est le plus facile ou le plus approprié dans le problème considéré). C'est-à-dire que

$$P[A] = \int_a^b \int_{g_1(x)}^{g_2(x)} f_{X,Y}(x,y)dydx = \int_a^b \left\{ \int_{g_1(x)}^{g_2(x)} f_{X,Y}(x,y)dy \right\} dx \qquad (1.4)$$

ou

$$P[A] = \int_c^d \int_{h_1(y)}^{h_2(y)} f_{X,Y}(x,y)dxdy = \int_c^d \left\{ \int_{h_1(y)}^{h_2(y)} f_{X,Y}(x,y)dx \right\} dy \qquad (1.5)$$

Remarques.
(i) Si les fonctions $g_1(x)$ et $g_2(x)$ (ou $h_1(y)$ et $h_2(y)$) sont constantes, et si la fonction $f_{X,Y}(x,y)$ peut être écrite comme suit:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

alors l'intégrale double qui donne $P[A]$ peut être exprimée comme un produit d'intégrales simples:

$$P[A] = \int_a^b f_X(x)dx \int_\alpha^\beta f_Y(y)dy$$

où $\alpha = g_1(x)\ \forall x$ et $\beta = g_2(x)\ \forall x$.

(ii) Réciproquement, on peut écrire le produit de deux intégrales simples comme une intégrale double:

$$\int_a^b f(x)dx \int_c^d g(y)dy = \int_a^b \int_c^d f(x)g(y)dydx$$

Exemple 1.3.5. Considérons la fonction

$$f_{X,Y}(x,y) = \begin{cases} e^{-x-y} \text{ si } x \geq 0 \text{ et } y \geq 0 \\ 0 \quad \text{ailleurs} \end{cases}$$

(voir la figure 1.1). Pour obtenir la probabilité que (X,Y) prenne une valeur

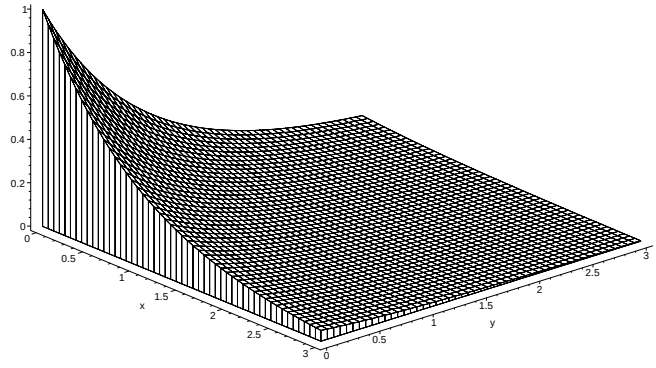

Fig. 1.1. Fonction de densité conjointe dans l'exemple 1.3.5

dans la région

$$A := \{(x,y) \in \mathbb{R}^2 \colon x \geq 0, y \geq 0, 0 \leq 2x + y \leq 1\}$$

(voir la figure 1.2), on calcule l'intégrale double

$$\begin{aligned} P[A] &= \int_0^{1/2} \int_0^{1-2x} e^{-x-y}\, dydx = \int_0^{1/2} \left\{ -e^{-x-y} \Big|_{y=0}^{y=1-2x} \right\} dx \\ &= \int_0^{1/2} \left\{ e^{-x} - e^{x-1} \right\} dx = -e^{-x} - e^{x-1} \Big|_0^{1/2} \\ &= -e^{-1/2} - e^{-1/2} + 1 + e^{-1} \simeq 0{,}1548 \end{aligned}$$

On peut aussi écrire que

$$\begin{aligned} P[A] &= \int_0^1 \int_0^{(1-y)/2} e^{-x-y}\,dxdy = \int_0^1 \left\{ -e^{-x-y}\Big|_{x=0}^{x=(1-y)/2} \right\} dy \\ &= \int_0^1 \left\{ e^{-y} - e^{-(1+y)/2} \right\} dy = -e^{-y} + 2e^{-(1+y)/2}\Big|_0^1 \\ &= -e^{-1} + 2e^{-1} + 1 - 2e^{-1/2} \simeq 0{,}1548 \end{aligned}$$

Remarque. Dans cet exemple, les fonctions $g_i(x)$ et $h_i(y)$, $i = 1, 2$, qui apparaissent dans les équations (1.4) et (1.5) sont données par

$$g_1(x) \equiv 0, \quad g_2(x) = 1 - 2x, \quad h_1(y) \equiv 0 \quad \text{et} \quad h_2(y) = (1-y)/2$$

Si B est le rectangle défini par

$$B = \{(x, y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le 2\}$$

alors on a:

$$\begin{aligned} P[B] &= \int_0^2 \int_0^1 e^{-x-y}\,dxdy = \int_0^1 e^{-x}\,dx \int_0^2 e^{-y}\,dy \\ &= \left(-e^{-x}\Big|_0^1 \right) \left(-e^{-y}\Big|_0^2 \right) = (1 - e^{-1})(1 - e^{-2}) \simeq 0{,}5466 \end{aligned}$$

♢

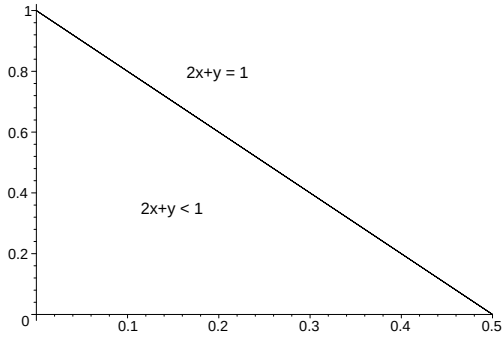

Fig. 1.2. Région d'intégration dans l'exemple 1.3.5

1.4 Séries infinies

Soit $a_1, a_2, \ldots, a_n, \ldots$ une *suite (infinie)* de nombres réels, où a_n est donné par une certaine formule ou règle, par exemple

$$a_n = \frac{1}{n+1} \quad \text{pour } n = 1, 2, \ldots$$

On symbolise la suite par $\{a_n\}_{n=1}^{\infty}$ ou simplement par $\{a_n\}$. On dit que la suite est *convergente* si $\lim_{n\to\infty} a_n$ existe; autrement, elle est *divergente*.

À partir de la suite $\{a_n\}_{n=1}^{\infty}$, on définit une nouvelle suite infinie par

$$S_1 = a_1, \quad S_2 = a_1 + a_2, \quad \ldots, \quad S_n = a_1 + a_2 + \cdots + a_n, \quad \ldots$$

Définition 1.4.1. *La suite infinie $S_1, S_2, \ldots, S_n, \ldots$ est représentée par $\sum_{n=1}^{\infty} a_n$ et est appelée* **série infinie**. *De plus, la somme finie $S_n := \sum_{k=1}^{n} a_k$ est appelée n^e* **somme partielle** *de la série. Finalement, si la limite $\lim_{n\to\infty} S_n$ existe (respectivement n'existe pas), on dit que la série est* **convergente** *(respectivement* **divergente***).*

En probabilités, l'ensemble des valeurs possibles d'une *variable aléatoire discrète* X peut être fini ou *infini dénombrable* (voir la page 35). Dans le deuxième cas, la *fonction de probabilité* p_X de X est telle que

$$\sum_{k=1}^{\infty} p_X(x_k) = 1$$

où $x_1, x_2, \ldots$ sont les valeurs possibles de X. Dans les cas les plus importants, les valeurs possibles de X sont en fait les entiers $0, 1, \ldots$

1.4.1 Séries géométriques

Un type particulier de série infinie que nous verrons au chapitre 3 est connu sous le nom de *série géométrique*. Ces séries sont de la forme

$$\sum_{n=1}^{\infty} ar^{n-1} \quad \text{ou} \quad \sum_{n=0}^{\infty} ar^n \tag{1.6}$$

où a et r sont des constantes réelles.

Proposition 1.4.1. *Si* $|r| < 1$*, alors la série géométrique* $S(a,r) := \sum_{n=0}^{\infty} ar^n$ converge *vers* $a/(1-r)$*. Si* $|r| \geq 1$ *(et* $a \neq 0$*), alors la série est* divergente.

Pour prouver les résultats ci-dessus, il suffit de considérer la n^{e} somme partielle de la série:

$$S_n = \sum_{k=1}^{n} ar^{k-1} = a + ar + ar^2 + \cdots + ar^{n-1}$$

On a:

$$rS_n = ar + ar^2 + ar^3 + \cdots + ar^{n-1} + ar^n$$

de sorte que

$$S_n - rS_n = a - ar^n \quad \stackrel{r\neq 1}{\Longrightarrow} \quad S_n = \frac{a(1-r^n)}{1-r}$$

De là, on déduit que

$$S := \lim_{n\to\infty} S_n = \begin{cases} \dfrac{a}{1-r} & \text{si } |r| < 1 \\ \text{n'existe pas} & \text{si } |r| > 1 \end{cases}$$

Si $r = 1$, alors $S_n = na$, de sorte que la série diverge (si $a \neq 0$). Finalement, on peut montrer que la série est aussi divergente si $r = -1$.

D'autres formules utiles associées aux séries géométriques sont les suivantes: si $|r| < 1$, alors

$$\sum_{k=1}^{\infty} ar^k = \frac{ar}{1-r}$$

et

$$\sum_{k=0}^{\infty} ak\, r^k = \frac{ar}{(1-r)^2} \tag{1.7}$$

En probabilités, on se sert aussi des *séries entières*, c'est-à-dire des séries de la forme

$$S(x) := a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n + \cdots$$

où a_k est une constante, pour $k = 0, 1, \ldots$ En particulier, on utilise le fait qu'il est possible d'exprimer des fonctions, par exemple la fonction exponentielle e^{cx}, sous la forme d'une série entière:

$$e^{cx} = 1 + cx + \frac{c^2}{2!}x^2 + \cdots + \frac{c^n}{n!}x^n + \cdots \quad \text{pour tout } x \in \mathbb{R} \tag{1.8}$$

Cette série est appelée *développement en série (entière)* de e^{cx}.

Remarque. Notons qu'une série géométrique $S(a, r)$ est une série entière $S(r)$ pour laquelle toutes les constantes a_k sont égales à a.

En général, une série entière converge pour toutes les valeurs de x dans un intervalle autour de 0. Si un certain développement en série est valable pour $|x| < R$ (> 0), on dit que R est le *rayon de convergence* de la série. Pour $|x| < R$, la série peut être dérivée et intégrée terme à terme:

$$S'(x) = a_1 + 2a_2x + \cdots + na_nx^{n-1} + \cdots \tag{1.9}$$

et

$$\int_0^x S(t)\,dt = a_0x + a_1\frac{x^2}{2} + \cdots + a_n\frac{x^{n+1}}{n+1} + \cdots$$

L'*intervalle de convergence* d'une série entière ayant un rayon de convergence $R > 0$ est au moins l'intervalle $(-R, R)$. La série peut converger ou ne pas converger pour $x = -R$ et $x = R$. Puisque

$$S(0) = a_0 \in \mathbb{R}$$

toute série entière converge pour $x = 0$. Si la série ne converge pas pour n'importe quel $x \neq 0$, on écrit que $R = 0$. Réciproquement, si la série converge pour tout $x \in \mathbb{R}$, alors $R = \infty$.

Pour calculer le rayon de convergence d'une série entière, on peut faire appel au *test du quotient d'Alembert*: supposons que la limite

$$L := \lim_{n\to\infty} \left|\frac{u_{n+1}}{u_n}\right| \tag{1.10}$$

existe. Alors la série $\sum_{n=0}^{\infty} u_n$

(a) converge *absolument* si $L < 1$;

(b) diverge si $L > 1$.

Remarques.

(i) Si la limite L dans la formule (1.10) n'existe pas, ou si $L = 1$, alors le test est non concluant. Il existe d'autres critères qui peuvent être utilisés, par exemple le critère de Raabe.

(ii) Une série infinie $\sum_{n=0}^{\infty} u_n$ converge *absolument* si $\sum_{n=0}^{\infty} |u_n|$ converge. Une série qui converge absolument est aussi convergente.

(iii) Dans le cas d'une série entière, on calcule

$$\lim_{n\to\infty}\left|\frac{a_{n+1}x^{n+1}}{a_n x^n}\right| = \lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right||x|$$

Par exemple, le développement en série de la fonction exponentielle e^{cx} donné dans la formule (1.8) est valable pour tout $x \in \mathbb{R}$. En effet, $a_n = c^n/n!$, de sorte que

$$\lim_{n\to\infty}\left|\frac{a_{n+1}x^{n+1}}{a_n x^n}\right| = \lim_{n\to\infty}\left|\frac{c}{n+1}\right||x| = 0 < 1 \quad \forall x \in \mathbb{R}$$

Exemple 1.4.1. Pour obtenir la *moyenne* d'une *variable aléatoire géométrique*, on peut calculer la somme infinie

$$\sum_{k=1}^{\infty} k(1-p)^{k-1}p = \frac{p}{1-p}\sum_{k=0}^{\infty} k(1-p)^k \overset{(1.7)}{=} \frac{p}{1-p}\frac{1-p}{p^2} = \frac{1}{p}$$

où $0 < p < 1$. Pour prouver la formule (1.7), on peut utiliser l'équation (1.9). ♢

Exemple 1.4.2. Une *variable aléatoire de Poisson* est telle que sa *fonction de probabilité* est donnée par

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{pour } x = 0, 1, \ldots$$

où λ est une constante positive. On a:

$$\sum_{x=0}^{\infty} p_X(x) = \sum_{x=0}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} \overset{(1.8)}{=} e^{-\lambda}e^{\lambda} = 1$$

comme requis. ♢

Exemple 1.4.3. La série entière

$$S_k(x) := 1 + kx + \frac{k(k-1)}{2!}x^2 + \cdots + \frac{k(k-1)(k-2)\cdots(k-n+1)}{n!}x^n + \cdots$$

est le développement en série de la fonction $(1+x)^k$ et est appelée *série binomiale*. En probabilités, k sera un entier naturel. Il s'ensuit que la série possède en fait un nombre fini de termes, et converge donc pour tout $x \in \mathbb{R}$. De plus, on peut écrire que

$$\frac{k(k-1)(k-2)\cdots(k-n+1)}{n!} = \frac{k!}{(k-n)!n!}$$

pour $n = 1, \ldots, k$. ♢

Pour conclure cette révision du calcul différentiel et intégral, nous donnons les principales formules logarithmiques:

$$\ln ab = \ln a + \ln b; \quad \ln a/b = \ln a - \ln b; \quad \ln a^b = b \ln a$$

On a aussi:

$$\ln e^{cx} = e^{\ln(cx)} = cx$$

et

$$e^{f(x)\ln x} = x^{f(x)}$$

1.5 Exercices du chapitre 1

Exercices résolus

Question n^o 1

Calculer $\lim_{x\downarrow 0} x \sin(1/x)$.

Solution. Lorsque x décroît vers 0, $1/x$ croît vers ∞. La fonction $\sin x$ ne converge pas lorsque x tend vers ∞. Cependant, parce que $-1 \leq \sin x \leq 1$, pour tout nombre réel x, on peut conclure que $\lim_{x\downarrow 0} x \sin(1/x) = 0$. Ce résultat peut être prouvé à partir de la définition de la limite d'une fonction. On a:

$$0 < |x| < \epsilon \quad \Longrightarrow \quad |x \sin(1/x)| \leq |x| < \epsilon$$

De là, on peut prendre $\delta = \epsilon$ dans la définition 1.1.1, et on peut en fait écrire que

$$\lim_{x\to 0} x \sin(1/x) = 0$$

Notons que la fonction $f(x) := x \sin(1/x)$ n'est pas définie en $x = 0$.

Question n^o 2

Pour quelles valeurs de x la fonction

$$f(x) = \begin{cases} \dfrac{\sin x}{x} & \text{si } x \neq 0 \\ \\ 1 & \text{si } x = 0 \end{cases}$$

est-elle continue?

Solution. Les fonctions $f_1(x) := \sin x$ et $f_2(x) := x$ sont continues, pour n'importe quel nombre réel x. De plus, on peut montrer que si $f_1(x)$ et $f_2(x)$ sont continues, alors $g(x) := f_1(x)/f_2(x)$ est aussi une fonction continue, pour tout x tel que $f_2(x) \neq 0$. Par conséquent, on peut affirmer que $f(x)$ est continue en n'importe quel $x \neq 0$.

Ensuite, en utilisant le développement en série entière de la fonction $\sin x$:

$$\sin x = x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 - \frac{1}{7!}x^7 + \cdots$$

on peut écrire que

$$\frac{\sin x}{x} = 1 - \frac{1}{3!}x^2 + \frac{1}{5!}x^4 - \frac{1}{7!}x^6 + \cdots$$

De là, on déduit que

$$\lim_{x\to 0} \frac{\sin x}{x} = 1$$

Remarque. Pour obtenir le résultat précédent, on peut aussi se servir de la règle de l'Hospital:

$$\lim_{x\to 0} \frac{\sin x}{x} = \lim_{x\to 0} \frac{\cos x}{1} = 1$$

Puisque, par définition

$$f(0) = 1 = \lim_{x\to 0} \frac{\sin x}{x}$$

on conclut que la fonction $f(x)$ est continue en tout $x \in \mathbb{R}$.

Question n$^{\text{o}}$ 3

Dériver la fonction $f(x) = \sqrt{3x+1}(2x^2+1)^2$.

Solution. Soit $g(x) = \sqrt{3x+1}$ et $h(x) = (2x^2+1)^2$. On a, en utilisant la *règle de dérivation en chaîne*:

$$g'(x) = \frac{1}{2\sqrt{3x+1}}(3) \quad \text{et} \quad h'(x) = 2(2x^2+1)(4x)$$

Donc,

$$f'(x) = g'(x)h(x) + g(x)h'(x) = \frac{3}{2\sqrt{3x+1}}(2x^2+1)^2 + \sqrt{3x+1}(8x)(2x^2+1)$$

$$\iff \quad f'(x) = (2x^2+1)\left\{\frac{3(2x^2+1)}{2\sqrt{3x+1}} + 8x\sqrt{3x+1}\right\}$$

Question n$^{\text{o}}$ 4

Trouver la limite $\lim_{x\downarrow 0} x \ln x$.

Solution. On a:

$$\lim_{x\downarrow 0} x \ln x = \lim_{x\downarrow 0} x \lim_{x\downarrow 0} \ln x = 0 \times (-\infty)$$

ce qui est indéterminé. En écrivant que

$$x \ln x = \frac{\ln x}{1/x}$$

on obtient que

$$\lim_{x\downarrow 0} x \ln x = \lim_{x\downarrow 0} \frac{\ln x}{1/x} = \frac{-\infty}{\infty}$$

On peut alors utiliser la règle de l'Hospital:

$$\lim_{x\downarrow 0} \frac{\ln x}{1/x} = \lim_{x\downarrow 0} \frac{1/x}{-1/x^2} = \lim_{x\downarrow 0} -x = 0$$

Question n$^{\text{o}}$ 5

Évaluer l'intégrale définie

$$I_5 := \int_1^e \frac{\ln x}{x} dx$$

Solution. Étant donné que la dérivée de $\ln x$ est $1/x$, on peut écrire que

$$I_5 := \int_1^e \frac{\ln x}{x} dx = \frac{(\ln x)^2}{2}\bigg|_1^e = \frac{(\ln e)^2 - (\ln 1)^2}{2} = \frac{1}{2}$$

Ce résultat peut être vérifié à l'aide de la technique d'intégration par substitution: en posant $y = \ln x \Leftrightarrow x = e^y$, on déduit que

$$\int \frac{\ln x}{x} dx = \int \frac{y}{e^y}(e^y)' dy = \int y dy = \frac{1}{2}y^2$$

Il s'ensuit que

$$I_5 = \int_0^1 y dy = \frac{1}{2} y^2 \Big|_0^1 = \frac{1}{2}$$

Question n^o 6

Trouver la valeur de l'intégrale définie

$$I_6 := \int_{-\infty}^{\infty} x^3 e^{-x^2/2} dx$$

Solution. On utilise la technique d'intégration par parties. On pose

$$u = x^2 \quad \text{et} \quad dv = x e^{-x^2/2} dx$$

Puisque $v = -e^{-x^2/2}$, il s'ensuit que

$$I_6 = -x^2 e^{-x^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 2x e^{-x^2/2} dx$$

Par la règle de l'Hospital, le terme constant ci-dessus est égal à 0; de plus, l'intégrale est donnée par

$$\int_{-\infty}^{\infty} 2x e^{-x^2/2} dx = -2e^{-x^2/2} \Big|_{-\infty}^{\infty} = 0$$

De là, on a que $I_6 = 0$.

Remarque. En probabilités, on déduit de ce résultat que l'*espérance mathématique* ou la *moyenne* du cube d'une *variable aléatoire gaussienne centrée réduite* est égale à zéro.

Question n^o 7

Trouver la transformée de Fourier de la fonction

$$f(x) = c e^{-cx} \quad \text{pour } x \geq 0$$

où c est une constante positive.

Solution. On a:

$$F(\omega) = \int_0^{\infty} e^{j\omega x} c e^{-cx} dx = c \int_0^{\infty} e^{(j\omega - c)x} dx = \frac{c}{j\omega - c} e^{(j\omega - c)x} \Big|_0^{\infty} = \frac{c}{c - j\omega}$$

Remarques.
(i) Le fait que j soit un *nombre (purement) imaginaire* ne cause aucun problème, car c'est une constante.

(ii) En probabilités, la fonction $F(\omega)$ obtenue ci-dessus est la *fonction caractéristique* de la *variable aléatoire* ayant une *distribution exponentielle* de *paramètre c*.

Question n° 8

Soit

$$f(x,y) = \begin{cases} x+y \text{ si } 0 \le x \le 1, 0 \le y \le 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

Calculer

$$I_8 := \int\int_A f(x,y)\,dxdy$$

où $A := \{(x,y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le 1, x^2 < y\}$ (voir la figure 1.3).

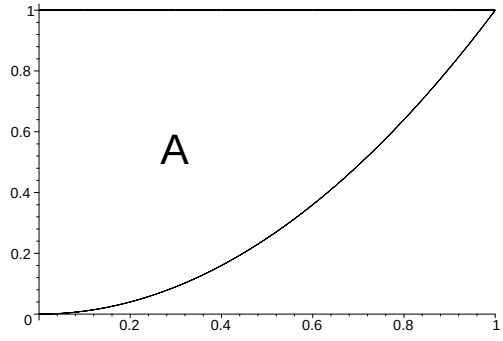

Fig. 1.3. Région A dans l'exercice résolu n° 8

Solution. On peut écrire que

$$I_8 = \int_0^1 \int_0^{\sqrt{y}} (x+y)dxdy = \int_0^1 \left(\frac{x^2}{2}\Big|_0^{\sqrt{y}} + y\sqrt{y} \right) dy$$
$$= \int_0^1 \left(\frac{y}{2} + y^{3/2} \right) dy = \frac{y^2}{4} + \frac{y^{5/2}}{5/2}\Big|_0^1 = \frac{1}{4} + \frac{2}{5} = \frac{13}{20}$$

ou que

$$\begin{aligned} I_8 &= \int_0^1 \int_{x^2}^1 (x+y)\,dy dx = \int_0^1 \left\{ x(1-x^2) + \frac{y^2}{2}\Big|_{x^2}^1 \right\} dx \\ &= \int_0^1 \left\{ x(1-x^2) + \frac{1}{2} - \frac{x^4}{2} \right\} dx = \frac{x^2}{2} - \frac{x^4}{4} + \frac{x}{2} - \frac{x^5}{10}\Big|_0^1 \\ &= \frac{1}{2} - \frac{1}{4} + \frac{1}{2} - \frac{1}{10} = \frac{13}{20} \end{aligned}$$

Remarque. On trouve facilement que

$$\int_0^1 \int_0^1 (x+y)\,dx dy = 1$$

De là, si $B := \{(x,y) \in \mathbb{R}^2 : 0 \le x \le 1, 0 \le y \le 1, x < y\}$, alors on peut écrire (par symétrie) que

$$\int\limits_B \int (x+y)\,dx dy = \frac{1}{2}$$

Question n$^{\text{o}}$ 9

Trouver la valeur de la série infinie

$$S_9 := \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots$$

Solution. Soit la série géométrique

$$S(1/2, 1/2) := \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots = \sum_{n=1}^{\infty} (1/2)^n$$

Cette série converge vers 1. De là, on peut écrire que

$$S_9 = 1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

ou que

$$S_9 = \frac{1}{4} S(1/2, 1/2) = \frac{1}{4}$$

Remarque. La somme S_9 représente la probabilité que le nombre de lancers nécessaires pour obtenir “face” avec une pièce de monnaie *bien équilibrée* soit supérieur à deux.

Question n$^{\text{o}}$ 10

Calculer

$$S_{10} := \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p$$

où $0 < p < 1$.

Solution. On a:

$$\sum_{k=1}^{\infty}(1-p)^{k-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$$

On peut dériver cette série terme à terme (deux fois). On obtient que

$$\frac{2}{p^3} = \frac{d^2}{dp^2}\sum_{k=1}^{\infty}(1-p)^{k-1} = \sum_{k=1}^{\infty}(k-1)(k-2)(1-p)^{k-3}$$

Puisque

$$\sum_{k=1}^{\infty}(k-1)(k-2)(1-p)^{k-3} = \sum_{k=3}^{\infty}(k-1)(k-2)(1-p)^{k-3}$$
$$= \sum_{n=1}^{\infty}(n+1)(n)(1-p)^{n-1} = \sum_{n=1}^{\infty} n^2(1-p)^{n-1} + \sum_{n=1}^{\infty} n(1-p)^{n-1}$$

on déduit (voir l'exemple 1.4.1) que

$$S_{10} = p\left(\frac{2}{p^3} - \frac{1}{p^2}\right) = \frac{2-p}{p^2}$$

Remarque. La somme calculée ci-dessus est la *moyenne* du carré d'une *variable aléatoire géométrique.*

Exercices

Question n$^{\text{o}}$ 1

Soit

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1/2 & \text{si } x = 0 \\ 1-(1/2)e^{-x} & \text{si } x > 0 \end{cases}$$

Calculer (a) $\lim_{x\uparrow 0} F(x)$, (b) $\lim_{x\downarrow 0} F(x)$ et (c) $\lim_{x\to 0} F(x)$.

Question n° 2

On considère la fonction

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1/3 & \text{si } 0 \leq x < 1 \\ 2/3 & \text{si } 1 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

Pour quelles valeurs de x la fonction $F(x)$ est-elle continue à gauche? continue à droite? continue?

Question n° 3

Trouver la limite lorsque x tend vers 0 de la fonction

$$f(x) = \frac{x^2 \sin(1/x)}{\sin x} \quad \text{pour } x \in \mathbb{R}$$

Question n° 4

La fonction

$$f(x) = \begin{cases} e^{1/x} & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$$

est-elle continue ou discontinue en $x = 0$? Justifier la réponse.

Question n° 5

Trouver la dérivée quatrième de la fonction $F(\omega) = e^{-\omega^2/2}$, pour n'importe quel $\omega \in \mathbb{R}$, et évaluer la dérivée en $\omega = 0$.

Question n° 6

Trouver la limite suivante:

$$\lim_{\epsilon \downarrow 0} \frac{\left(1 + x - \frac{\epsilon}{2}\right) e^{-x+(\epsilon/2)} - \left(1 + x + \frac{\epsilon}{2}\right) e^{-x-(\epsilon/2)}}{\epsilon}$$

Question n° 7

Déterminer la dérivée seconde de $f(x) = \sqrt[3]{2x^2}$.

Question n° 8

Calculer la dérivée de

$$f(x) = 1 + \sqrt[3]{x^2} \quad \text{pour } x \in \mathbb{R}$$

et trouver la valeur de x qui minimise cette fonction.

Question n$^{\text{o}}$ 9

Utiliser le fait que

$$\int_0^\infty x^{n-1}e^{-x}dx = (n-1)! \quad \text{pour } n = 1, 2, \ldots$$

pour évaluer l'intégrale

$$\int_0^\infty x^{n-1}e^{-cx}dx$$

où c est une constante positive.

Question n$^{\text{o}}$ 10

Utiliser la formule suivante:

$$\int_{-\infty}^\infty e^{-(x-m)^2/2}dx = \sqrt{2\pi}$$

où m est une constante réelle, pour calculer l'intégrale définie

$$\int_{-\infty}^\infty x^2 e^{-x^2/2}dx$$

Question n$^{\text{o}}$ 11

Évaluer l'intégrale impropre

$$\int_0^\infty \frac{e^{-x}}{x}dx$$

Question n$^{\text{o}}$ 12

Trouver une primitive de la fonction

$$f(x) = e^{-x}\sin x \quad \text{pour } x \in \mathbb{R}$$

Question n$^{\text{o}}$ 13

Trouver la transformée de Fourier de la fonction

$$f(x) = \frac{c}{2}e^{-c|x|} \quad \text{pour } x \in \mathbb{R}$$

où c est une constante positive.

Question n° 14

Soit

$$f(x,y) = \begin{cases} \frac{3}{2}x \text{ si } 1 \le x \le 2, 1 \le y \le 2, x \le y \\ 0 \quad \text{ailleurs} \end{cases}$$

Calculer l'intégrale double définie

$$\int_A \int f(x,y)\,dxdy$$

où

$$A := \{(x,y) \in \mathbb{R}^2 \colon 1 \le x \le 2, 1 \le y \le 2, x^2 < y\}$$

Question n° 15

La *convolution* de deux fonctions, f et g, est notée $f * g$ et est définie par

$$(f * g)(x) = \int_{-\infty}^{\infty} f(y)g(x-y)\,dy$$

Soit

$$f(x) = \begin{cases} ce^{-cx} \text{ si } x \ge 0 \\ 0 \quad \text{si } x < 0 \end{cases}$$

où c est une constante positive. Supposons que $g(x) \equiv f(x)$ (c'est-à-dire que g est *identique* à f). Trouver $(f * g)(x)$.

Remarque. En probabilités, la convolution de f_X et f_Y est la *fonction de densité (de probabilité)* de la somme $Z := X+Y$ des *variables aléatoires indépendantes* X et Y. Le résultat de l'exercice ci-dessus implique que la somme de deux variables aléatoires qui présentent une *distribution exponentielle* de même *paramètre* c, et sont indépendantes, présente une *distribution gamma* de paramètres $\alpha = 2$ et $\lambda = c$.

Question n° 16

Montrer que

$$I := \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1$$

en écrivant d'abord que

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-z^2/2} dz \int_{-\infty}^{\infty} e^{-w^2/2} dw = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(z^2+w^2)/2} dz\,dw$$

puis en utilisant les *coordonnées polaires*. C'est-à-dire qu'on pose $z = r\cos\theta$ et $w = r\sin\theta$ (avec $r \geq 0$), de sorte que

$$r = \sqrt{z^2 + w^2} \quad \text{et} \quad \theta = \text{tg}^{-1}(w/z)$$

Remarque. On trouve que $I^2 = 1$. Justifier que cela implique que $I = 1$ (et non pas $I = -1$).

Question n° 17

Déterminer la valeur de la série infinie

$$\frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \cdots + \frac{(-1)^n}{n!}x^n + \cdots$$

Question n° 18

Soit

$$S(q) = \sum_{n=1}^{\infty} q^{n-1}$$

où $0 < q < 1$. Calculer

$$\int_0^{1/2} S(q)dq$$

Question n° 19

(a) Calculer la série infinie

$$M(t) := \sum_{k=0}^{\infty} e^{tk} e^{-\alpha} \frac{\alpha^k}{k!}$$

où $\alpha > 0$.

(b) Évaluer la dérivée seconde $M''(t)$ en $t = 0$.

Remarque. La fonction $M(t)$ est la *fonction génératrice des moments* de la *variable aléatoire* X qui présente une *distribution de Poisson* de *paramètre* α. De plus, $M''(0)$ donne la *moyenne* de X^2.

Question n° 20

(a) Déterminer la valeur de la série entière

$$G(z) := \sum_{k=0}^{\infty} z^k (1-p)^k p$$

où $z \in \mathbb{R}$ et $p \in (0,1)$ sont tels que $|z| < (1-p)^{-1}$.

(b) Calculer

$$\frac{d^k}{dz^k} G(z) \quad \text{pour } k = 0, 1, \ldots$$

en $z = 0$.

Remarque. La fonction $G_X(z) := \sum_{k=0}^{\infty} z^k p_X(k)$ est appelée *fonction génératrice* de la *variable aléatoire discrète* X prenant ses valeurs dans l'ensemble $\{0, 1, \ldots\}$ et possédant la *fonction de probabilité* $p_X(k)$. De plus,

$$\frac{1}{k!} \frac{d^k}{dz^k} G_X(z)\bigg|_{z=0} = p_X(k)$$

Questions à choix multiple

Question n° 1

Calculer la limite

$$\lim_{x \to 1} \frac{\sqrt{x} - 1}{1 - x}$$

(a) -1 (b) $-1/2$ (c) $1/2$ (d) 1 (e) n'existe pas

Question n° 2

Soit $u(x)$ la fonction de Heaviside (voir la page 3). On définit la fonction $h(x) = e^{u(x)/2}$ pour $x \in \mathbb{R}$. Calculer la limite de $h(x)$ lorsque x tend vers 0.

(a) $1/2$ (b) 1 (c) $(1 + e^{1/2})/2$ (d) $e^{1/2}$ (e) n'existe pas

Question n° 3

Évaluer la dérivée seconde de la fonction

$$F(\omega) = \frac{2(2 + j\omega)}{(4 + \omega^2)} \quad \text{pour } \omega \in \mathbb{R}$$

en $\omega = 0$.

(a) $-1/4$ (b) $-1/2$ (c) 0 (d) $1/4$ (e) $1/2$

Question n° 4

Trouver la limite suivante:

$$\lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x$$

Indication. Prendre d'abord le logarithme naturel de l'expression, et ensuite utiliser le fait que $\ln x$ est une fonction continue.

(a) 0 (b) 1 (c) e (d) ∞ (e) n'existe pas

Question n$^{\text{o}}$ 5

Utiliser la formule

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}b} \exp\left\{ -\frac{(x-a)^2}{2b^2} \right\} dx = 1$$

où a et b (> 0) sont des constantes, pour évaluer l'intégrale définie

$$\int_0^{\infty} \frac{1}{\sqrt{2\pi}} e^{-2x^2} dx$$

(a) 1/4 (b) 1/2 (c) 1 (d) 2 (e) 4

Question n$^{\text{o}}$ 6

Calculer l'intégrale définie

$$\int_0^1 x \ln x dx$$

(a) $-\infty$ (b) -1 (c) $-1/2$ (d) $-1/4$ (e) 0

Question n$^{\text{o}}$ 7

On suppose que

$$f(x) = \begin{cases} 1 \text{ si } 0 \leq x \leq 1 \\ 0 \text{ ailleurs} \end{cases}$$

et que $g(x) = f(x)$ pour tout $x \in \mathbb{R}$. Trouver $(f * g)(3/2)$, où $*$ symbolise la *convolution* de f et g (voir l'exercice n$^{\text{o}}$ 15, page 30).

(a) 0 (b) 1/2 (c) 1 (d) 3/2 (e) 2

Question n$^{\text{o}}$ 8

Soit

$$f(x,y) = \begin{cases} 2 - x - y \text{ si } 0 < x < 1, 0 < y < 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

Calculer l'intégrale double suivante:

$$\int_A \int f(x,y)\, dxdy$$

où

$$A := \{(x, y) \in \mathbb{R}^2 : 0 < x < 1, 0 < y < 1, x + y > 1\}$$

(a) 1/6 (b) 1/4 (c) 1/3 (d) 1/2 (e) 5/6

Question n^o 9

Trouver la valeur du logarithme naturel du *produit infini*

$$\prod_{n=1}^{\infty} e^{1/2^n} = e^{1/2} \times e^{1/4} \times e^{1/8} \times \cdots$$

(a) -1 (b) $-1/2$ (c) 0 (d) 1/2 (e) 1

Question n^o 10

On définit $a_0 = 0$ et

$$a_k = e^{-c} \frac{c^{|k|}}{2|k|!} \quad \text{pour } k = \ldots, -2, -1, 1, 2, \ldots$$

où $c > 0$ est une constante. Trouver la valeur de la série infinie $\sum_{k=-\infty}^{\infty} a_k$.

(a) $1 - e^{-c}$ (b) 0 (c) 1 (d) $(1/2)e^{-c}$ (e) $1 + e^{-c}$

2

Probabilités élémentaires

Ce premier chapitre consacré aux probabilités contient les définitions et concepts de base, tels qu'ils peuvent être enseignés dans un cours de niveau collégial, c'est-à-dire préuniversitaire, sur ce sujet. Cependant, la gamme de problèmes qui peuvent être composés en utilisant les formules des *probabilités élémentaires* est très étendue, particulièrement en *analyse combinatoire*, et il est nécessaire de faire beaucoup d'exercices pour bien assimiler ces notions de base.

2.1 Expériences aléatoires

Une **expérience aléatoire** est une expérience qui, au moins théoriquement, peut être répétée aussi souvent que l'on veut et dont on ne peut pas prédire le résultat; par exemple, le jet d'un dé à jouer. Chaque fois que l'expérience est répétée, un **résultat élémentaire** est obtenu. L'ensemble des résultats élémentaires d'une expérience aléatoire est appelé **espace échantillon**; on le notera Ω.

Les espaces échantillons peuvent être discrets ou continus.

(a) Espaces échantillons discrets:

(i) si le nombre de résultats possibles est fini. Par exemple, on lance un dé et l'on observe le nombre qui apparaît; alors $\Omega = \{1, 2, \ldots, 6\}$.

(ii) Si le nombre de résultats possibles est *infini dénombrable*; cela signifie qu'il y a un nombre infini de résultats possibles mais qu'on peut associer un nombre naturel à chaque résultat. Par exemple, on lance un dé jusqu'à ce qu'on obtienne le chiffre "6" et on compte le nombre de lancers effectués avant d'obtenir ce premier "6"; alors on a: $\Omega = \{0, 1, 2, \ldots\}$. Cet ensemble est équivalent à

l'ensemble des entiers naturels $\{1, 2, \ldots\}$ puisque l'on peut associer le nombre naturel $k+1$ à tout élément $k = 0, 1, \ldots$ de Ω.

(b) Espaces échantillons continus: si l'espace échantillon contient un ou plusieurs intervalles; l'espace échantillon est alors *infini non dénombrable*. Par exemple, on lance un dé jusqu'à ce que l'on obtienne un "6" et l'on calcule le temps que cela a pris pour obtenir ce premier "6". Dans ce cas, on a: $\Omega = \{t \in \mathbb{R} : t > 0\}$ (ou $\Omega = (0, \infty)$).

2.2 Événements

Définition 2.2.1. *Un* **événement** *est un ensemble de résultats élémentaires. C'est-à-dire que c'est un sous-ensemble de l'espace échantillon Ω. En particulier, chaque résultat élémentaire est un événement, de même que l'espace échantillon lui-même.*

Remarques.
(i) Un résultat élémentaire est parfois appelé événement *simple*, tandis qu'un événement *composé* est constitué d'au moins deux résultats élémentaires.

(ii) On désignera les événements par A, B, C, etc.

Définition 2.2.2. *On dit que deux événements, A et B, sont* **incompatibles** *(ou* **mutuellement exclusifs***) si leur intersection est vide. On écrit alors: $A \cap B = \emptyset$.*

Exemple 2.2.1. Considérons l'expérience qui consiste à lancer un dé à jouer et à observer le nombre qui apparaît. On a: $\Omega = \{1, 2, 3, 4, 5, 6\}$. Soit les événements

$$A = \{1, 2, 4\}, \quad B = \{2, 4, 6\} \quad \text{et} \quad C = \{3, 5\}$$

On a:

$$A \cup B = \{1, 2, 4, 6\}, \quad A \cap B = \{2, 4\} \quad \text{et} \quad A \cap C = \emptyset$$

Donc, A et C sont des événements incompatibles. De plus, on peut écrire que $A' = \{3, 5, 6\}$, où le symbole $'$ désigne le **complément** de l'événement. ♢

Pour représenter un espace échantillon et des événements, on utilise souvent un **diagramme de Venn** comme dans la figure 2.1. En général, pour trois événements on a le diagramme de la figure 2.2, à la page 37.

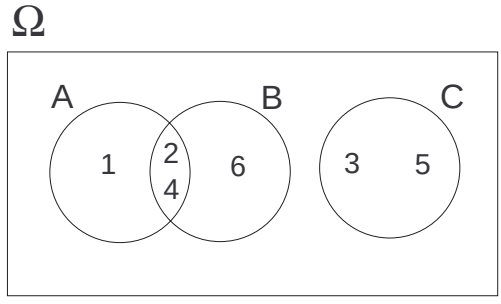

Fig. 2.1. Diagramme de Venn pour l'exemple 2.2.1

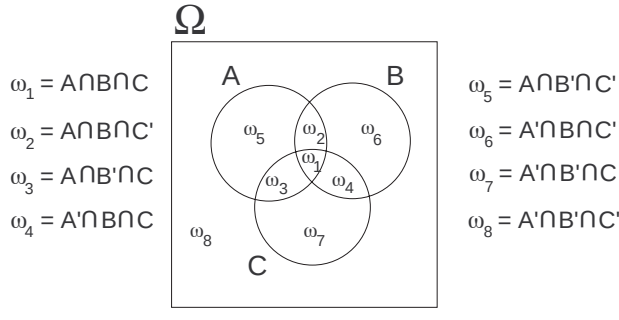

Fig. 2.2. Diagramme de Venn pour trois événements quelconques

2.3 Probabilité

Définition 2.3.1. *La* **probabilité** *d'un événement $A \subseteq \Omega$, notée $P[A]$, est un nombre réel obtenu en appliquant à A la fonction P qui possède les propriétés suivantes:*
(i) $0 \leq P[A] \leq 1$*;*
(ii) si $A = \Omega$*, alors* $P[A] = 1$*;*
(iii) si $A = A_1 \cup A_2 \cup \ldots \cup A_n$*, où* $A_1, \ldots, A_n$ *sont des événements incompatibles, alors on peut écrire que*

$$P[A] = \sum_{i=1}^{n} P[A_i] \quad \textit{pour } n = 2, 3, \ldots, \infty \tag{2.1}$$

Remarques.
(i) En fait, il suffit de poser que $P[A] \geq 0$ dans la définition, car on peut montrer que

$$P[A] + P[A'] = 1 \tag{2.2}$$

ce qui implique que $P[A] = 1 - P[A'] \leq 1$.
(ii) On a aussi les résultats suivants:

$$P[\emptyset] = 0 \quad \text{et} \quad P[A] \leq P[B] \quad \text{si } A \subset B \tag{2.3}$$

(iii) La définition de la probabilité d'un événement est motivée par la notion de **fréquence relative**. Par exemple, supposons que l'on effectue un grand nombre de fois l'expérience aléatoire qui consiste à lancer un dé, et que l'on désire obtenir la probabilité d'un des résultats possibles de cette expérience, soit les entiers $1, 2, \ldots, 6$. La *fréquence relative* du résultat k est la quantité $f_k(n)$ définie par

$$f_k(n) = \frac{N_k(n)}{n} \tag{2.4}$$

où $N_k(n)$ est le nombre de fois que le résultat possible k s'est produit au cours de n lancers du dé. On peut écrire que

$$0 \leq f_k(n) \leq 1 \quad \text{pour } k = 1, 2, \ldots, 6 \tag{2.5}$$

et que

$$\sum_{k=1}^{6} f_k(n) = 1 \tag{2.6}$$

En effet, on a évidemment: $N_k(n) \in \{0, 1, \ldots, n\}$, de sorte que $f_k(n) \in [0, 1]$, et

$$\sum_{k=1}^{6} f_k(n) = \frac{N_1(n) + \ldots + N_6(n)}{n} = \frac{n}{n} = 1 \tag{2.7}$$

De plus, si A est un événement qui contient deux résultats possibles, par exemple "1" et "2", alors

$$f_A(n) = f_1(n) + f_2(n) \tag{2.8}$$

car les résultats 1 et 2 ne peuvent pas se produire lors du *même* lancer du dé.

Finalement, la probabilité du résultat k peut théoriquement être obtenue en prenant la limite de $f_k(n)$ lorsque le nombre n de lancers tend vers l'infini:

$$P[\{k\}] = \lim_{n \to \infty} f_k(n) \tag{2.9}$$

Puisque la probabilité d'un événement peut être exprimée en fonction de la fréquence relative de cet événement, il est naturel que les propriétés que les

probabilités possèdent soit plus ou moins calquées sur celle des fréquences relatives.

Parfois, la probabilité d'un résultat élémentaire est simplement égale à 1 divisé par le nombre total de résultats élémentaires. Dans ce cas, on dit que les résultats élémentaires sont **équiprobables**. Par exemple, si on lance un dé *bien équilibré* (ou *non truqué*), alors on a: $P[\{1\}] = P[\{2\}] = \ldots = P[\{6\}] = 1/6$.

Si les résultats élémentaires r_i ne sont *pas* équiprobables, on peut (essayer de) se servir de la formule suivante:

$$P[A] = \sum_{r_i \in A} P[\{r_i\}] \tag{2.10}$$

Toutefois, cette formule n'est utile que si l'on connaît la probabilité de chacun des résultats élémentaires r_i qui constituent l'événement A.

Maintenant, si A et B sont des événements incompatibles, alors on déduit de la troisième propriété de $P[\cdot]$ que $P[A \cup B] = P[A] + P[B]$. Si A et B ne sont pas incompatibles, on peut montrer (voir la figure 2.3) que

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \tag{2.11}$$

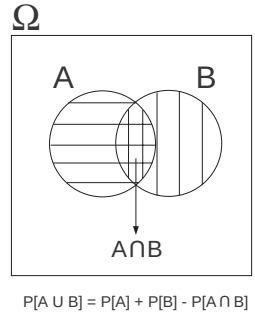

Fig. 2.3. Probabilité de l'union de deux événements quelconques

De même, dans le cas de trois événements quelconques, on a:

$$P[A \cup B \cup C] = P[A] + P[B] + P[C] - P[A \cap B] - P[A \cap C] - P[B \cap C] + P[A \cap B \cap C] \tag{2.12}$$

Exemple 2.3.1. Les trois options les plus populaires pour un certain modèle de voiture neuve sont: A: la transmission automatique, B: le moteur V6 et C: le climatiseur. À partir des ventes précédentes, on peut supposer que $P[A] = 0{,}70$, $P[B] = 0{,}75$, $P[C] = 0{,}80$, $P[A \cup B] = 0{,}80$, $P[A \cup C] = 0{,}85$, $P[B \cup C] = 0{,}90$ et $P[A \cup B \cup C] = 0{,}95$, où $P[A]$ désigne la probabilité qu'un acheteur choisisse l'option A, etc. Calculer la probabilité de chacun des événements suivants:
(a) l'acheteur choisit au moins une des trois options;
(b) l'acheteur ne choisit aucune des trois options;
(c) l'acheteur choisit uniquement le climatiseur;
(d) l'acheteur choisit exactement une des trois options.

Solution. (a) On cherche $P[A \cup B \cup C] = 0{,}95$ (par hypothèse).

(b) On cherche $P[A' \cap B' \cap C'] = 1 - P[A \cup B \cup C] = 1$ - 0,95 = 0,05.

(c) L'événement dont on nous demande la probabilité est $A' \cap B' \cap C$. On peut écrire que

$$P[A' \cap B' \cap C] = P[A \cup B \cup C] - P[A \cup B] = 0{,}95 - 0{,}8 = 0{,}15$$

(d) On veut maintenant calculer

$$\begin{aligned} &P[(A \cap B' \cap C') \cup (A' \cap B \cap C') \cup (A' \cap B' \cap C)] \\ &\overset{\text{inc.}}{=} P[A \cap B' \cap C'] + P[A' \cap B \cap C'] + P[A' \cap B' \cap C] \\ &= 3P[A \cup B \cup C] - P[A \cup B] - P[A \cup C] - P[B \cup C] \\ &= 3(0{,}95) - 0{,}8 - 0{,}85 - 0{,}9 = 0{,}3 \end{aligned}$$

Remarques.
(i) La mention de "inc." au-dessus du signe "=" signifie que l'égalité est vraie à cause de l'*incompatibilité* des événements. Nous allons utiliser cette notation à plusieurs reprises dans ce manuel pour justifier le passage d'une expression à une autre.

(ii) La probabilité de chacun des huit résultats élémentaires est indiquée dans le diagramme de la figure 2.4. On calcule d'abord

$$P[A \cap B] = P[A] + P[B] - P[A \cup B] = 0{,}7 + 0{,}75 - 0{,}8 = 0{,}65$$

De même, on a:

$$\begin{aligned} P[A \cap C] &= 0{,}7 + 0{,}8 - 0{,}85 = 0{,}65 \\ P[B \cap C] &= 0{,}75 + 0{,}8 - 0{,}9 = 0{,}65 \\ P[A \cap B \cap C] &= P[A \cup B \cup C] - P[A] - P[B] - P[C] \\ &\quad + P[A \cap B] + P[A \cap C] + P[B \cap C] \\ &= 0{,}95 - 0{,}7 - 0{,}75 - 0{,}8 + 3(0{,}65) = 0{,}65 \end{aligned}$$

♢

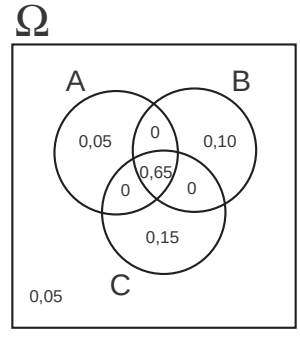

Fig. 2.4. Diagramme de Venn pour l'exemple 2.3.1

2.4 Probabilité conditionnelle

Définition 2.4.1. *La* **probabilité conditionnelle** *de l'événement A, étant donné que l'événement B s'est produit, est notée $P[A \mid B]$ et est définie par (voir la figure 2.5)*

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} \quad si\ P[B] > 0 \tag{2.13}$$

À partir de la définition ci-dessus, on obtient la **règle de multiplication**:

$$P[A \cap B] = P[A]P[B \mid A] \quad \text{si } P[A] > 0 \tag{2.14}$$

et

$$P[A \cap B] = P[B]P[A \mid B] \quad \text{si } P[B] > 0 \tag{2.15}$$

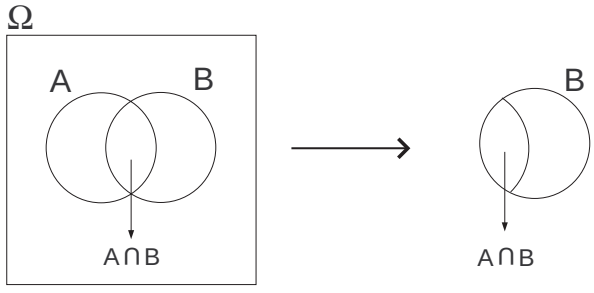

Fig. 2.5. Notion de probabilité conditionnelle

Définition 2.4.2. *Soit A et B deux événements tels que $P[A]P[B] > 0$. On dit que A et B sont des événements* **indépendants** *si*

$$P[A \mid B] = P[A] \quad ou \quad P[B \mid A] = P[B] \tag{2.16}$$

On déduit de la règle de multiplication que A et B sont indépendants si et seulement si (ssi)

$$P[A \cap B] = P[A]P[B] \tag{2.17}$$

En fait, cette équation est la définition de l'indépendance des événements A et B dans le cas général où l'on peut avoir: $P[A]P[B] = 0$. Cependant, la définition 2.4.2 est plus intuitive, tandis que la définition générale de l'indépendance donnée par la formule (2.17) est purement mathématique.

En général, les événements $A_1, A_2, \ldots, A_n$ sont indépendants ssi

$$P[A_{i_1} \cap \ldots \cap A_{i_k}] = \prod_{j=1}^{k} P[A_{i_j}] \tag{2.18}$$

pour $k = 2, 3, \ldots, n$, où $A_{i_l} \neq A_{i_m}$ si $l \neq m$.

Remarque. Si A et B sont deux événements incompatibles, alors ils ne peuvent *pas* être indépendants, à moins que $P[A]P[B] = 0$. En effet, dans le cas où $P[A]P[B] > 0$, on a:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{P[\emptyset]}{P[B]} = \frac{0}{P[B]} = 0 \neq P[A]$$

Exemple 2.4.1. Un appareil est constitué de deux composants, A et B, susceptibles de tomber en panne. Les composants sont placés en parallèle et ne sont pas

indépendants (voir la figure 2.6). On estime à 0,2 la probabilité d'une panne du composant A, à 0,8 la probabilité d'une panne du composant B si le composant A est en panne, et à 0,4 la probabilité d'une panne du composant B si le composant A n'est pas en panne.

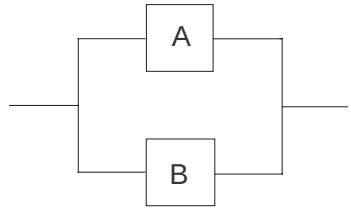

Fig. 2.6. Système pour la partie (a) de l'exemple 2.4.1

(a) Calculer la probabilité d'une panne (i) du composant B; (ii) de l'appareil.

Solution. Soit A (respectivement B): le composant A (respectivement B) est en panne. Par hypothèse, on peut écrire que $P[A] = 0{,}2$, $P[B \mid A] = 0{,}8$ et $P[B \mid A'] = 0{,}4$.

(i) On peut écrire (voir la figure 2.7) que

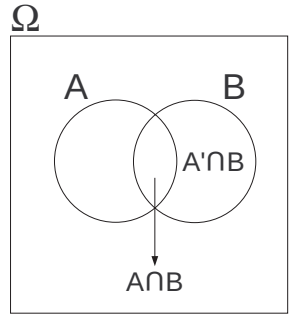

Fig. 2.7. Diagramme de Venn pour la partie (a) de l'exemple 2.4.1

$$\begin{aligned} P[B] &= P[A \cap B] + P[A' \cap B] = P[B \mid A]P[A] + P[B \mid A']P[A'] \\ &= (0{,}8)(0{,}2) + (0{,}4)(0{,}8) = 0{,}48 \end{aligned}$$

(ii) On cherche $P[\text{Panne de l'appareil}] = P[A \cap B] = P[B \mid A]P[A] = 0{,}16$.

(b) Afin d'augmenter la fiabilité de l'appareil, on installe un troisième composant, C, de telle sorte que les composants A, B et C sont placés en parallèle (voir la figure 2.8). La probabilité que le composant C tombe en panne est de 0,2, et ce, indépendamment de l'état (en panne ou non) des composants A et B. Calculer la probabilité que l'appareil formé des composants A, B et C tombe en panne.

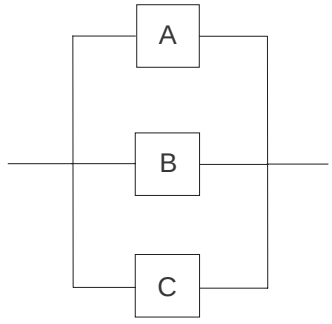

Fig. 2.8. Système pour la partie (b) de l'exemple 2.4.1

Solution. On a: $P[C] = 0{,}2$ et C est indépendant de A et de B. On cherche

$$P[A \cap B \cap C] \overset{\text{ind.}}{=} P[A \cap B]P[C] \overset{\text{(a)(ii)}}{=} (0{,}16)(0{,}2) = 0{,}032$$

♢

2.5 Probabilité totale

Soit $B_1, B_2, \ldots, B_n$ des événements *incompatibles* et *exhaustifs*; c'est-à-dire que l'on a:

$$B_i \cap B_j = \emptyset \quad \text{si } i \neq j \quad \text{et} \quad \bigcup_{i=1}^{n} B_i = \Omega$$

On dit que les événements B_i constituent une **partition** de l'espace échantillon Ω. Il s'ensuit que

$$P\left[\bigcup_{i=1}^{n} B_i\right] = \sum_{i=1}^{n} P[B_i] = P[\Omega] = 1$$

Soit maintenant A un événement quelconque. On peut écrire (voir la figure 2.9) que

$$P[A] = \sum_{i=1}^{n} P[A \cap B_i] = \sum_{i=1}^{n} P[A \mid B_i]P[B_i] \tag{2.19}$$

(la deuxième égalité ci-dessus étant valide lorsque $P[B_i] > 0$ pour $i = 1, 2, \ldots, n$).

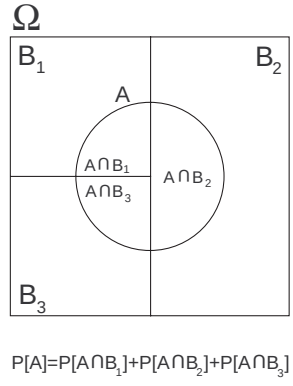

Fig. 2.9. Exemple de la règle de la probabilité totale avec $n = 3$

Remarque. Cette formule est parfois appelée **règle de la probabilité totale**.

Finalement, supposons que l'on désire calculer $P[B_i \mid A]$ pour $i = 1, \ldots, n$. On a:

$$P[B_i \mid A] = \frac{P[B_i \cap A]}{P[A]} = \frac{P[A \mid B_i]P[B_i]}{\sum_{j=1}^{n} P[A \cap B_j]} = \frac{P[A \mid B_i]P[B_i]}{\sum_{j=1}^{n} P[A \mid B_j]P[B_j]} \tag{2.20}$$

Cette formule est appelée **formule de Bayes**.

Remarque. On a aussi (la *règle de Bayes*):

$$P[B \mid A] = \frac{P[A \mid B]P[B]}{P[A]} \quad \text{si } P[A]P[B] > 0 \tag{2.21}$$

Exemple 2.5.1. Les machines M_1, M_2 et M_3 produisent respectivement 500, 1000 et 1500 pièces par jour, dont 5 %, 6 % et 7 % sont défectueuses. Une pièce produite par l'une des trois machines est prise au hasard, à la fin d'une journée donnée, et elle est défectueuse. Quelle est la probabilité qu'elle ait été produite par la machine M_3?

Solution. Soit A_i: la pièce prise au hasard a été produite par la machine M_i, pour $i = 1, 2, 3$, et soit D: la pièce prise au hasard est défectueuse. On cherche

$$P[A_3 \mid D] = \frac{P[D \mid A_3]P[A_3]}{\sum_{i=1}^{3} P[D \mid A_i]P[A_i]} = \frac{(0,07)\left(\frac{1500}{3000}\right)}{(0{,}05)\left(\frac{1}{6}\right) + (0{,}06)\left(\frac{1}{3}\right) + (0{,}07)\left(\frac{1}{2}\right)}$$
$$= \frac{105}{190} \simeq 0{,}5526$$

♢

2.6 Analyse combinatoire

Supposons que l'on effectue une expérience aléatoire que l'on peut diviser en deux étapes. Lors de la première étape, le résultat A_1 *ou* le résultat A_2 peut se produire; lors de la seconde étape, l'un ou l'autre des résultats B_1, B_2 ou B_3 peut se produire. On peut utiliser un **arbre** pour décrire l'espace échantillon de cette expérience aléatoire comme dans la figure 2.10.

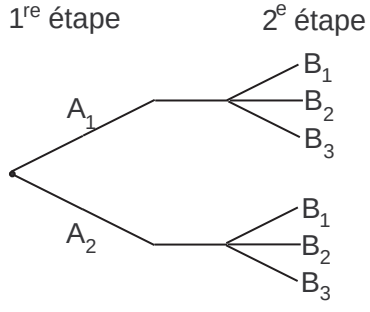

Fig. 2.10. Exemple d'arbre

Exemple 2.6.1. Des essais avec un nouvel alcootest ont permis d'établir que i) 5 fois sur 100 l'alcootest s'est révélé positif alors que la personne soumise au test n'était pas en état d'ébriété; (ii) 90 fois sur 100 l'alcootest s'est révélé positif alors que la personne soumise au test était réellement en état d'ébriété. De plus, on estime que 1 % des personnes soumises au test sont réellement en état d'ébriété.

Calculer la probabilité
(a) que l'alcootest soit positif pour la prochaine personne soumise au test;
(b) qu'une personne soit en état d'ébriété étant donné que l'alcootest est positif.

Solution. Soit A: l'alcootest est positif, et E: la personne soumise au test est en état d'ébriété. À partir des hypothèses ci-dessus, on peut construire l'arbre de la figure 2.11, où l'on a indiqué sur les branches les probabilités *marginales* des événements E et E', de même que les probabilités conditionnelles des événements A et A', étant donné E ou E'. De plus, on sait par la règle de multiplication que le produit de ces probabilités égale la probabilité des intersections $E \cap A$, etc.

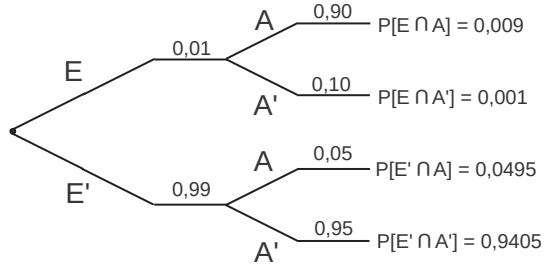

Fig. 2.11. Arbre dans l'exemple 2.6.1

(a) On a:

$$P[A] = P[E \cap A] + P[E' \cap A] = 0{,}0585$$

(b) On calcule

$$P[E \mid A] = \frac{P[E \cap A]}{P[A]} \overset{\text{(a)}}{=} \frac{0{,}009}{0{,}0585} \simeq 0{,}1538$$

Notons que cette probabilité est très faible. Si l'on suppose que 60 % des personnes soumises au test sont en état d'ébriété (plutôt que 1 %), alors on trouve que $P[A] = 0{,}56$ et $P[E \mid A] \simeq 0{,}9643$, ce qui est beaucoup plus raisonnable. Cet alcootest est donc efficace seulement si on le fait passer à des gens qu'on soupçonne d'être en état d'ébriété. ♢

Remarque. En général, si une expérience aléatoire est constituée de k étapes et s'il y a n_j résultats possibles lors de la j^{e} étape, pour $j = 1, \ldots, k$, alors il y a $n_1 \times \ldots \times n_k$ résultats élémentaires dans l'espace échantillon. C'est le **principe de multiplication**.

Supposons maintenant que l'on dispose de n objets *distincts* et que l'on prend, au hasard et *sans* remise, r objets parmi les n, où $r \in \{0, \ldots, n\}$. Le nombre d'*arrangements* possibles est donné par

$$n \times (n-1) \times \ldots \times [n-(r-1)] = \frac{n!}{(n-r)!} := P_r^n \tag{2.22}$$

Le symbole P_r^n désigne le nombre de **permutations** de n objets distincts pris r à la fois. L'*ordre* des objets est important.

Rappel. On a: $n! = 1 \times 2 \times \ldots \times n$ pour $n = 1, 2, 3, \ldots$ et $0! = 1$, par définition.

Exemple 2.6.2. Si l'on dispose de quatre lettres différentes (par exemple, a, b, c et d), alors on peut former

$$P_3^4 = \frac{4!}{(4-3)!} = \frac{4!}{1!} = 24$$

"mots" de trois lettres différents si l'on utilise chaque lettre au plus une fois. On peut utiliser un arbre pour en établir la liste. ♢

Finalement, si l'*ordre* des objets n'a pas d'importance, alors le nombre de façons de prendre, au hasard et *sans* remise, r objets parmi n objets distincts est donné par

$$\frac{n \times (n-1) \times \ldots \times [n-(r-1)]}{r!} = \frac{n!}{r!(n-r)!} := C_r^n \equiv \binom{n}{r} \tag{2.23}$$

Le symbole C_r^n, ou $\binom{n}{r}$, désigne le nombre de **combinaisons** de n objets distincts pris r à la fois.

Remarques.
(i) Chaque combinaison de r objets permet d'obtenir $r!$ permutations différentes, car

$$P_r^r = \frac{r!}{(r-r)!} = \frac{r!}{0!} = r!$$

(ii) De plus, il est facile de vérifier que $C_r^n = C_{n-r}^n$.

Exemple 2.6.3. On prend 3 pièces, au hasard et *sans* remise, parmi 10 pièces, dont 2 sont défectueuses. Quelle est la probabilité que l'on obtienne au moins une pièce défectueuse?

Solution. Soit F: au moins une pièce est défectueuse parmi les trois pièces prises au hasard. On peut écrire que

$$P[F] = 1 - P[F'] = 1 - \frac{C_0^2 \cdot C_3^8}{C_3^{10}}$$
$$= 1 - \frac{1 \cdot \frac{8!}{3!5!}}{\frac{10!}{3!7!}} = 1 - \frac{6 \times 7}{9 \times 10} = \frac{8}{15} = 0{,}5\bar{3}$$

♢

2.7 Exercices du chapitre 2

Exercices résolus

Question n$^{\text{o}}$ 1

On considère l'expérience aléatoire suivante: on lance un dé bien équilibré; si (et seulement si) on obtient un "6", on lance le dé une deuxième fois. Combien y a-t-il de résultats élémentaires dans l'espace échantillon Ω?

Solution. On a: $\Omega = \{1, 2, 3, 4, 5, (6, 1), \ldots, (6, 6)\}$; ainsi, il y a $5 + 6 = 11$ résultats élémentaires.

Question n$^{\text{o}}$ 2

Soit $\Omega = \{e_1, e_2, e_3\}$, où $P[\{e_i\}] > 0$ pour $i = 1, 2, 3$. Combien de partitions différentes de Ω, en excluant la partition $\emptyset$, Ω peut-on former?

Solution. On peut former quatre partitions différentes de Ω: $\{e_1\}, \{e_2, e_3\}$, ou $\{e_2\}, \{e_1, e_3\}$, ou $\{e_3\}, \{e_1, e_2\}$, ou $\{e_1\}, \{e_2\}, \{e_3\}$.

Question n$^{\text{o}}$ 3

On lance un dé bien équilibré deux fois de façon indépendante; sachant que l'on a obtenu un nombre pair lors du premier lancer, quelle est la probabilité que la somme des deux nombres obtenus égale 4?

Solution. Soit S_i: la somme des deux nombres obtenus égale i et $D_{j,k}$: le nombre obtenu lors du j^{e} lancer est k. On cherche

$$P[S_4 \mid D_{1,2} \cup D_{1,4} \cup D_{1,6}] = \frac{P[D_{1,2} \cap D_{2,2}]}{P[D_{1,2} \cup D_{1,4} \cup D_{1,6}]} \overset{\text{ind.}}{=} \frac{(1/6)^2}{1/2} = 1/18$$

Question n° 4

Supposons que $P[A] = P[B] = 1/4$ et que $P[A \mid B] = P[B]$. Calculer $P[A \cap B']$.

Solution. On a: $P[A \mid B] = P[B] = P[A] = 1/4 \Rightarrow A$ et B sont des événements indépendants. Il s'ensuit que

$$P[A \cap B'] = P[A]P[B'] = (1/4)(3/4) = 3/16$$

Question n° 5

Un système est constitué de trois composants indépendants. Le système fonctionne si au moins deux des trois composants fonctionnent. Si la fiabilité de chaque composant est égale à 0,95, quelle est la fiabilité du système?

Solution. Soit F_i: le composant i fonctionne et F_S: le système fonctionne. Par symétrie et incompatibilité, on peut écrire que

$$\begin{aligned} P[F_S] &= 3 \times P[F_1 \cap F_2 \cap F_3'] + P[F_1 \cap F_2 \cap F_3] \\ &\overset{\text{ind.}}{=} 3 \times (0{,}95)(0{,}95)(0{,}05) + (0{,}95)^3 = 0{,}99275 \end{aligned}$$

Question n° 6

Supposons que $P[A \cap B] = 1/4$, $P[A \mid B'] = 1/8$ et $P[B] = 1/2$. Calculer $P[A]$.

Solution. On a:

$$\begin{aligned} P[A] &= P[A \cap B] + P[A \cap B'] = \frac{1}{4} + P[A \mid B']P[B'] \\ &= \frac{1}{4} + \left(\frac{1}{8}\right) \cdot \left(1 - \frac{1}{2}\right) = \frac{5}{16} \end{aligned}$$

Question n° 7

Sachant que l'on a obtenu au moins une fois le résultat "face" lors de trois lancers indépendants d'une pièce de monnaie bien équilibrée, quelle est la probabilité que l'on ait obtenu exactement trois fois "face"?

Solution. Soit A_i: i "faces" ont été obtenues. On cherche

$$P[A_3 \mid A_1 \cup A_2 \cup A_3] = \frac{P[A_3]}{P[A_1 \cup A_2 \cup A_3]} \overset{\text{ind.}}{=} \frac{(1/2)^3}{1 - (1/2)^3} = \frac{1}{7}$$

Question n^{o} 8

Supposons que $P[B \mid A_1] = 1/2$ et que $P[B \mid A_2] = 1/4$, où A_1 et A_2 sont deux événements équiprobables formant une partition de Ω. Calculer $P[A_1 \mid B]$.

Solution. On a:

$$P[A_1 \mid B] = \frac{P[B \mid A_1]P[A_1]}{P[B \mid A_1]P[A_1] + P[B \mid A_2]P[A_2]} \overset{P[A_1]=P[A_2]}{=} \frac{1/2}{1/2 + 1/4} = \frac{2}{3}$$

Question n^{o} 9

Trois chevaux, a, b et c, font une course. Si le résultat bac signifie que b arrive le premier, a le deuxième et c le troisième, alors l'ensemble de tous les résultats possibles est

$$\Omega = \{abc, acb, bac, bca, cab, cba\}$$

On suppose que $P[\{abc\}] = P[\{acb\}] = 1/18$ et que les quatre autres résultats élémentaires ont une probabilité de $2/9$. De plus, on définit les événements

$$A\text{: } a \text{ précède } b \quad \text{et} \quad B\text{: } a \text{ précède } c$$

(a) Les événements A et B forment-ils une partition de Ω? Justifier votre réponse.

(b) Les événements A et B sont-ils indépendants? Justifier.

Solution. On a:

$$A = \{abc, acb, cab\} \quad \text{et} \quad B = \{abc, acb, bac\}$$

(a) Puisque $A \cap B = \{abc, acb\} \neq \emptyset$, A et B ne forment pas une partition de Ω. (De plus, $A \cup B \neq \Omega$.)

(b) $P[A] = \frac{1}{18} + \frac{1}{18} + \frac{2}{9} = P[B]$ et $P[A \cap B] = P[\{abc, acb\}] = \frac{2}{18}$. Puisque

$$P[A]P[B] = \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{9} = P[A \cap B]$$

on peut affirmer que les événements A et B sont indépendants.

Question n^{o} 10

Soit ε une expérience aléatoire pour laquelle il y a trois résultats élémentaires: A, B et C. Supposons que l'on répète ε indéfiniment de façon indépendante. Calculer, en fonction de $P[A]$ et $P[B]$, la probabilité que A se produise avant B.

Indications.
(i) Vous pouvez utiliser la règle de la probabilité totale.
(ii) Soit D: A se produit avant B. Alors on peut écrire que

$$P\left[D \mid C \text{ se produit au premier essai}\right] = P\left[D\right]$$

Solution. Soit A_1: A se produit au premier essai. De même pour B_1 et C_1. Alors on peut écrire que

$$\begin{aligned} & P[D] = P[D \mid A_1]P[A_1] + P[D \mid B_1]P[B_1] + P[D \mid C_1]P[C_1] \\ \iff \quad & P[D] = 1 \cdot P[A] + 0 + P[D]P[C] \\ \iff \quad & P[D] = \frac{P[A]}{1 - P[C]} = \frac{P[A]}{P[A] + P[B]} \end{aligned}$$

Question n$^{\text{o}}$ 11

Des transistors sont pris au hasard et sans remise dans une boîte qui en contient un très grand nombre, dont certains sont bons et d'autres sont défectueux, et testés un à la fois. On continue jusqu'à ce que l'on ait obtenu un transistor défectueux ou que trois transistors aient été testés. Décrire l'espace échantillon Ω pour cette expérience aléatoire.

Solution. Soit B: le transistor testé est bon et D: le transistor testé est défectueux. Alors on a: $\Omega = \{D, BD, BBD, BBB\}$.

Remarque. On pourrait aussi écrire que $\Omega = \{B', BB', BBB', BBB\}$.

Question n$^{\text{o}}$ 12

Soit A et B des événements tels que $P\left[A\right] = 1/3$ et $P\left[B' \mid A\right] = 5/7$. Calculer $P\left[B\right]$ si B est un sous-ensemble de A.

Solution. On a: $P[B \mid A] = 1 - P[B' \mid A] = 2/7$ et

$$P[B \mid A] = \frac{P[B \cap A]}{P[A]} = \frac{P[B]}{P[A]} \quad \text{car } B \subset A$$

Il s'ensuit que

$$P[B] = P[B \mid A]P[A] = \frac{2}{21} \simeq 0{,}0952$$

Question n^{o} 13

Dans une certaine université, la proportion de professeurs titulaires, agrégés, adjoints et de chargés de cours est de 30 %, 40 %, 20 % et 10 %, dont 60 %, 70 %, 90 % et 40 % respectivement ont un doctorat. Quelle est la probabilité qu'une personne prise au hasard parmi celles enseignant à cette université ait un doctorat?

Solution. Soit D: la personne a un doctorat, et T (respectivement Ag, Ad, CC): la personne est professeur titulaire (respectivement professeur agrégé, professeur adjoint, chargé de cours). On peut écrire que

$$\begin{aligned} P[D] &= P[D \mid T]P[T] + P[D \mid Ag]P[Ag] + P[D \mid Ad]P[Ad] + P[D \mid CC]P[CC] \\ &= (0{,}6)(0{,}3) + (0{,}7)(0{,}4) + (0{,}9)(0{,}2) + (0{,}4)(0{,}1) = 0{,}68 \end{aligned}$$

Question n^{o} 14

Tous les articles en inventaire dans un magasin ont un code constitué de cinq lettres. Si l'on n'utilise jamais deux fois la même lettre dans un code, combien de codes différents peut-il y avoir?

Solution. Le nombre de codes différents est donné par

$$26 \times 25 \times 24 \times 23 \times 22 = 7.893.600 \quad (= P_5^{26})$$

Remarque. Dans ce manuel, nous allons utiliser la convention qui veut que l'on sépare les nombres supérieurs ou égaux à dix mille en mettant un point entre les tranches de trois chiffres. Par exemple, on écrit: 1.000.000 pour représenter un million. On pourrait simplement laisser un espace entre ces tranches, mais la notation avec les points nous semble meilleure.

Question n^{o} 15

On lance un dé bien équilibré deux fois. Soit

A: le premier nombre qui apparaît est un "6";

B: la somme des deux nombres obtenus est égale à 7;

C: la somme des deux nombres obtenus est égale à 7 ou 11.

(a) Calculer $P[B \mid C]$.

(b) Calculer $P[A \mid B]$.

(c) Les événements A et B sont-ils indépendants? Justifier votre réponse.

Solution. On a: $\Omega = \{(1,1), (1,2), \ldots, (6,6)\}$. Il y a 36 résultats élémentaires *équiprobables*.

(a) $B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$ et $C = B \cup \{(5,6),(6,5)\}$. Alors

$$P[B \mid C] = \frac{P[B \cap C]}{P[C]} = \frac{P[B]}{P[C]} = \frac{6/36}{8/36} = 3/4$$

(b) On a:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} \overset{\text{(a)}}{=} \frac{P[\{(6,1)\}]}{6/36} = \frac{1/36}{6/36} = 1/6$$

(c) Puisque le dé est bien équilibré, on peut écrire que $P[A] = 1/6 \overset{\text{(b)}}{=} P[A \mid B]$. Donc, A et B sont des événements indépendants.

Question n$^{\text{o}}$ 16

Un banlieusard possède deux véhicules automobiles, dont l'un est une voiture compacte et l'autre une fourgonnette. Les trois quarts du temps, il utilise la voiture compacte pour se rendre au travail et le reste du temps il se sert de la fourgonnette. Quand il utilise la voiture compacte, il arrive à la maison avant 17 h 30 75 % du temps; lorsqu'il se sert de la fourgonnette, il arrive à la maison avant 17 h 30 60 % du temps (mais la fourgonnette possède un climatiseur). Calculer la probabilité

(a) qu'il arrive à la maison avant 17 h 30 lors d'une journée donnée;

(b) qu'il se soit servi de la voiture compacte s'il n'est pas arrivé avant 17 h 30;

(c) qu'il utilise la fourgonnette et qu'il arrive après 17 h 30;

(d) qu'il arrive à la maison avant 17 h 30 lors de deux journées consécutives (indépendantes) et qu'il n'utilise pas le même véhicule au cours de ces deux journées.

Solution. (a) Soit

A: le banlieusard arrive à la maison avant 17 h 30;

B: le banlieusard utilise la voiture compacte.

On cherche

$$\begin{aligned} P[A] &= P[A \mid B]P[B] + P[A \mid B']P[B'] \\ &= (0{,}75)(0{,}75) + (0{,}60)(0{,}25) = 0{,}7125 \end{aligned}$$

(b) On calcule

$$P[B \mid A'] = P[A' \mid B]\frac{P[B]}{P[A']} \overset{\text{(a)}}{=} (1 - 0{,}75)\frac{0{,}75}{1 - 0{,}7125} \simeq 0{,}6522$$

(c) On a: $P[A' \cap B'] = P[A' \mid B']P[B'] = (1 - 0{,}60)(0{,}25) = 0{,}1$.

(d) Par indépendance, on cherche

$$\begin{aligned} & P[A \cap B] \cdot P[A \cap B'] + P[A \cap B'] \cdot P[A \cap B] \\ &= 2P[A \mid B]P[B] \cdot P[A \mid B']P[B'] = 2(0{,}75)^2 \cdot (0{,}60)(0{,}25) \\ &= 0{,}16875 \end{aligned}$$

Question n° 17

On prévoit de la pluie la moitié du temps dans une certaine région pendant une période donnée. On estime que les prévisions de la météo sont exactes les deux tiers du temps. Monsieur X sort tous les jours et il craint beaucoup d'être pris sous la pluie sans parapluie. Par conséquent, il emporte toujours son parapluie si l'on prévoit de la pluie; de plus, il emporte même son parapluie le tiers du temps lorsqu'on ne prévoit pas de pluie. Calculer la probabilité qu'il pleuve et que Monsieur X n'ait pas son parapluie.

Solution. On définit les événements suivants: A: il pleut, B: on prévoyait de la pluie, et C: Monsieur X a son parapluie. On cherche

$$\begin{aligned} P[A \cap C'] &= P[A \cap \underbrace{B \cap C'}_{\emptyset}] + P[A \cap B' \cap C'] = P[A \cap B' \cap C'] \\ &= P[A \mid B' \cap C']P[B' \cap C'] = \frac{1}{3}P[C' \mid B']P[B'] \\ &= \frac{1}{3} \times \frac{2}{3} \times \frac{1}{2} = \frac{1}{9} \end{aligned}$$

Question n° 18

On lance un dé bien équilibré trois fois de façon indépendante. Soit F: le premier nombre obtenu est inférieur au deuxième nombre obtenu, qui est lui-même inférieur au troisième nombre obtenu. Calculer $P[F]$.

Solution. On a: $\Omega = \{(1,1,1), \ldots, (1,1,6), \ldots, (6,6,6)\}$. Il y a $6^3 = 216$ résultats élémentaires, qui sont équiprobables. Par *énumération*, on trouve qu'il y a 20 triplets pour lesquels l'événement F se réalise:

$$(1,2,3),\ (1,2,4),\ (1,2,5),\ (1,2,6),\ (1,3,4), \ldots$$

Par conséquent, la probabilité recherchée est de $\frac{20}{216} \simeq 0{,}0926$.

Question n$^{\text{o}}$ 19

On considère l'ensemble des familles ayant exactement deux enfants. On suppose que chacun des deux enfants a autant de chances d'être un garçon ou une fille. Soit

A: les deux sexes sont représentés chez les enfants;

B: au plus un enfant est une fille.

(a) Les événements A et B' sont-ils incompatibles? Justifier votre réponse.

(b) Les événements A et B sont-ils indépendants? Justifier votre réponse.

(c) On suppose aussi que la probabilité que le troisième enfant d'une famille soit un garçon est de $11/20$ si les deux premiers enfants sont des garçons, de $2/5$ si les deux premiers enfants sont des filles, et de $1/2$ dans les autres cas. Sachant que le troisième enfant d'une certaine famille est un garçon, quelle est la probabilité que les deux premiers soient aussi des garçons?

Solution. On a: $\Omega = \{(F,F),(F,G),(G,F),(G,G)\}$. De plus,

$$A = \{(F,G),(G,F)\} \quad \text{et} \quad B = \{(F,G),(G,F),(G,G)\}$$

(a) Puisque $B' = \{(F,F)\}$, on a: $A \cap B' = \emptyset$. Donc A et B' sont incompatibles.

(b) On peut écrire que $P[A] = 2/4$ et $P[B] = 3/4$. Étant donné que

$$P[A \cap B] = P[A] = 2/4 \neq P[A]P[B]$$

A et B ne sont pas indépendants.

(c) Soit G_i (respectivement F_i): le i^{e} enfant est un garçon (respectivement une fille). On calcule d'abord

$$\begin{aligned} P[G_3] &= P[G_3 \mid G_1 \cap G_2]P[G_1 \cap G_2] + P[G_3 \mid F_1 \cap F_2]P[F_1 \cap F_2] \\ &\quad + P[G_3 \mid (G_1 \cap F_2) \cup (F_1 \cap G_2)]P[(G_1 \cap F_2) \cup (F_1 \cap G_2)] \\ &= \left(\frac{11}{20}\right)\left(\frac{1}{4}\right) + \left(\frac{2}{5}\right)\left(\frac{1}{4}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{39}{80} \end{aligned}$$

Alors on peut écrire que

$$P[G_1 \cap G_2 \mid G_3] = \frac{P[G_3 \mid G_1 \cap G_2]P[G_1 \cap G_2]}{P[G_3]} = \frac{\frac{11}{20} \cdot \frac{1}{4}}{\frac{39}{80}} \simeq 0{,}2821$$

Exercices

Question n° 1

On étudie le trafic (dans un sens) sur deux routes, 1 et 2, qui se rejoignent pour former la route 3 (voir la figure 2.12). Les routes 1 et 2 ont la même capacité (nombre de voies) et la route 3 a une capacité plus grande que la route 1 et la route 2. Aux heures de pointe, la probabilité que le trafic soit excessif est de 0,1 sur la route 1 et de 0,3 sur la route 2. De plus, sachant qu'il est excessif sur la route 2, il l'est aussi sur la route 1 une fois sur trois. Posons:

A, B, C : le trafic est excessif sur les routes 1, 2, 3, respectivement

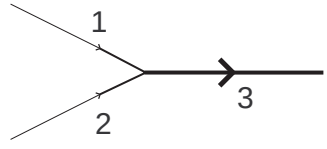

Fig. 2.12. Figure pour l'exercice n° 1

(a) Calculer la probabilité que le trafic soit excessif

(i) sur les routes 1 et 2;

(ii) sur la route 2 sachant qu'il l'est sur la route 1;

(iii) seulement sur la route 1;

(iv) seulement sur la route 2;

(v) sur la route 1 ou sur la route 2;

(vi) ni sur la route 1 ni sur la route 2.

(b) Sur la route 3, le trafic est excessif avec une probabilité de
1 s'il l'est sur la route 1 et sur la route 2;
0,15 s'il l'est sur la route 2 seulement;
0,1 s'il ne l'est pas ni sur la route 1 ni sur la route 2.
Calculer la probabilité que le trafic soit excessif

(i) sur la route 3;

(ii) sur la route 1 sachant qu'il est excessif sur la route 3.

Question n^{o} 2

On lance un dé, puis une pièce de monnaie; si l'on obtient "pile", alors on lance le dé une deuxième fois. On suppose que le dé et la pièce de monnaie sont bien équilibrés. Quelle est la probabilité de chacun des événements suivants:

(a) obtenir "face" ou un "6" la première fois qu'on lance le dé;

(b) n'obtenir aucun "6";

(c) obtenir exactement un "6";

(d) avoir obtenu "face", étant donné que l'on a obtenu exactement un "6".

Question n^{o} 3 (voir l'exemple 2.4.1)

Un appareil est formé de deux composants, A et B, susceptibles de tomber en panne. Les composants sont placés en parallèle et, ainsi, l'appareil est en panne seulement si les deux composants sont en panne. Les deux composants ne sont pas indépendants et on estime à

0,2 la probabilité d'une panne du composant A;

0,8 la probabilité d'une panne du composant B si le composant A est en panne;

0,4 la probabilité d'une panne du composant B si le composant A n'est pas en panne.

(a) Calculer la probabilité d'une panne

(i) du composant A si le composant B est en panne;

(ii) d'exactement un composant.

(b) Afin d'augmenter la fiabilité de l'appareil, on installe un troisième composant, C, de telle sorte que les composants A, B et C sont placés en parallèle. La probabilité que le composant C tombe en panne est de 0,2, et ce, indépendamment de l'état (en panne ou non) des composants A et B. Étant donné que l'appareil est en fonctionnement, quelle est la probabilité que le composant C soit en panne?

Question n^{o} 4

Dans une usine de fabrication de composants électroniques, on assure le contrôle de la qualité à l'aide de trois tests comme suit:

- chaque composant est assujetti au test n^{o} 1;
- si le composant réussit le test n^{o} 1, alors il est soumis au test n^{o} 2;
- si le composant réussit le test n^{o} 2, alors il est soumis au test n^{o} 3;
- dès qu'un composant échoue à l'un des tests, on le retourne pour réparation.

On définit les événements

$$A_i\text{: le composant échoue au test n}^{\text{o}}\ i, \quad \text{pour } i = 1, 2, 3$$

Par expérience, on estime que

$$P[A_1] = 0{,}1, \quad P\left[A_2 \mid A_1'\right] = 0{,}05, \quad P\left[A_3 \mid A_1' \cap A_2'\right] = 0{,}02$$

Les résultats élémentaires de l'espace échantillon Ω sont: $\omega_1 = A_1$, $\omega_2 = A_1' \cap A_2$, $\omega_3 = A_1' \cap A_2' \cap A_3$ et $\omega_4 = A_1' \cap A_2' \cap A_3'$.

(a) Calculer la probabilité de chacun des résultats élémentaires.

(b) Soit R l'événement: le composant doit être réparé.

(i) Exprimer R en fonction de A_1, A_2, A_3.

(ii) Calculer la probabilité de R.

(iii) Calculer $P[A_1' \cap A_2 \mid R]$.

(c) On teste trois composants et on définit les événements

$$R_k\text{: le } k^{\text{e}} \text{ composant doit être réparé}, \quad \text{pour } k = 1, 2, 3$$

et

$$B\text{: au moins un des trois composants réussit les trois tests}$$

On suppose que les événements R_k sont indépendants.

(i) Exprimer B en fonction de R_1, R_2, R_3.

(ii) Calculer $P[B]$.

Question n$^{\text{o}}$ 5

Soit A, B et C des événements tels que $P[A] = 1/2$, $P[B] = 1/3$, $P[C] = 1/4$ et $P[A \cap C] = 1/12$. De plus, A et B sont incompatibles. Calculer $P[A \mid B \cup C]$.

Question n$^{\text{o}}$ 6

Dans un groupe de 20.000 hommes et 10.000 femmes, 6 % des hommes et 3 % des femmes souffrent d'une certaine maladie. Quelle est la probabilité qu'un membre de ce groupe souffrant de la maladie en question soit un homme?

Question n$^{\text{o}}$ 7

Deux jetons sont tirés au hasard et sans remise d'une urne contenant dix jetons numérotés de 1 à 10. Quelle est la probabilité que le plus grand des deux nombres obtenus soit 3?

Question n^o 8

On considère le système dans la figure 2.13, à la page 60. Tous les composants tombent en panne indépendamment les uns des autres. Au cours d'une certaine période, les composants de type A tombent en panne avec une probabilité de 0,3 et le composant B (respectivement C) tombe en panne avec une probabilité de 0,01 (respectivement 0,1). Calculer la probabilité que le système ne soit pas en panne à la fin de cette période.

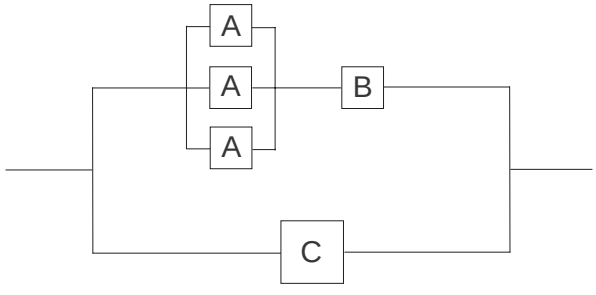

Fig. 2.13. Figure pour l'exercice n^o 8

Question n^o 9

On prélève (sans remise) un échantillon de taille 20 dans un lot de taille infinie contenant 2 % d'articles défectueux. Calculer la probabilité d'obtenir au moins un article défectueux dans l'échantillon.

Question n^o 10

Un lot contient 10 articles, dont 1 est défectueux. On examine les articles un par un, sans remise, jusqu'à ce que l'article défectueux soit découvert. Quelle est la probabilité que celui-ci soit (a) le deuxième article examiné? (b) le neuvième article examiné?

Question n^o 11

Un sac renferme deux pièces de monnaie: une pièce bien équilibrée et une autre avec laquelle on est certain d'obtenir "face". On tire une pièce au hasard et on la lance. Sachant que l'on a obtenu "face", calculer

(a) la probabilité que l'on ait tiré la pièce bien équilibrée;

(b) la probabilité que l'on obtienne "face" en lançant la pièce une deuxième fois.

Question n^{o} 12

Le diagnostic d'un médecin concernant l'un de ses patients est incertain. Il hésite entre trois maladies. Par expérience, on a réussi à établir les tableaux suivants:

S_i	S_1	S_2	S_3	S_4
$P[M_1 \mid S_i]$	0,2	0,1	0,6	0,4

S_i	S_1	S_2	S_3	S_4
$P[M_2 \mid S_i]$	0,2	0,5	0,5	0,3

S_i	S_1	S_2	S_3	S_4
$P[M_3 \mid S_i]$	0,6	0,3	0,1	0,2

où les M_i représentent les maladies et les S_i des symptômes. De plus, on suppose que les quatre symptômes sont incompatibles, exhaustifs et équiprobables.

(a) Indépendamment du symptôme qu'il présente, quelle est la probabilité que le patient souffre de la première maladie?

(b) Quelle est la probabilité que le patient souffre de la deuxième maladie et présente le symptôme S_1?

(c) Sachant que le patient souffre de la troisième maladie, quelle est la probabilité qu'il présente le symptôme S_2?

(d) Si l'on a deux patients indépendants, quelle est la probabilité qu'ils ne souffrent pas de la même maladie? On supposera que les trois maladies sont incompatibles.

Question n^{o} 13

On considère un système constitué de quatre composants fonctionnant indépendamment les uns des autres et reliés comme dans la figure 2.14.

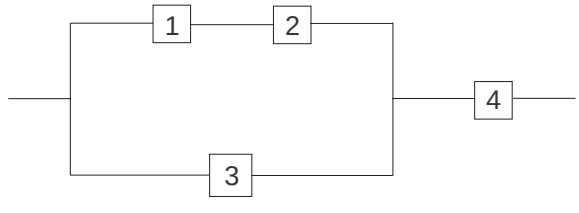

Fig. 2.14. Figure pour l'exercice n^{o} 13

La probabilité de fonctionnement de chaque composant est supposée constante, au cours d'une certaine période, et est donnée par le tableau suivant:

Composant	1	2	3	4
Fiabilité	0,9	0,95	0,95	0,99

(a) Quelle est la probabilité que le système fonctionne à la fin de cette période?

(b) Quelle est la probabilité que le composant n° 3 soit en panne et que le système fonctionne malgré tout?

(c) Quelle est la probabilité qu'au moins un des quatre composants soit en panne?

(d) Étant donné que le système est en panne, quelle est la probabilité que celui-ci se remette à fonctionner si l'on remplace le composant n° 1 par un composant identique (qui fonctionne)?

Question n° 14

Une boîte contient 20 transistors, dont 8 de marque A et 12 de marque B. Deux transistors sont tirés au hasard et sans remise. Quelle est la probabilité qu'ils ne soient pas de la même marque?

Question n° 15

Quelle est la fiabilité du système illustré dans la figure 2.15 si les quatre composants fonctionnent indépendamment les uns des autres et ont tous une fiabilité égale à 0,9 à un instant donné?

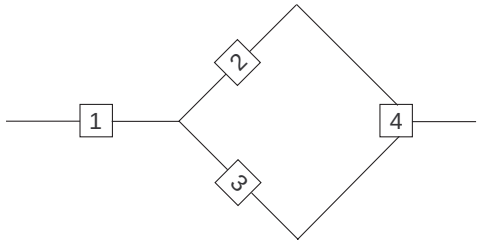

Fig. 2.15. Figure pour l'exercice n° 15

Question n° 16

Soit A_1 et A_2 deux événements tels que $P[A_1] = 1/4$, $P[A_1 \cap A_2] = 3/16$ et $P[A_2 \mid A_1'] = 1/8$. Calculer $P[A_2']$.

Question n° 17

On lance une pièce de monnaie bien équilibrée jusqu'à ce que l'on obtienne une "face" ou que le nombre total de lancers égale 3. Sachant que l'expérience

aléatoire s'est terminée avec une "face", quelle est la probabilité qu'on ait lancé la pièce une seule fois?

Question n^{o} 18

Dans une salle, il y a quatre étudiants de première année, six étudiantes de première année, six étudiants de deuxième année et des étudiantes de deuxième année. Combien doit-il y avoir d'étudiantes de deuxième année si l'on veut que le sexe et l'année soient indépendants lorsqu'une personne est prise au hasard dans la salle?

Question n^{o} 19

Les magasins M_1, M_2 et M_3 de la même entreprise emploient respectivement 50, 70 et 100 personnes dont 50 %, 60 % et 75 % sont des femmes. Une personne de cette entreprise est prise au hasard et c'est une femme. Quelle est la probabilité qu'elle travaille pour le magasin M_3?

Question n^{o} 20

Les oxydes d'azote nocifs constituent 20 %, par rapport au poids, de tous les polluants présents dans l'air dans une certaine région métropolitaine. Les échappements des automobiles sont responsables de 70 % de ces oxydes d'azote, mais de seulement 10 % de tous les autres polluants dans l'air. De la pollution totale dont les échappements d'automobiles sont responsables, quel pourcentage sont des oxydes d'azote nocifs?

Question n^{o} 21

Trois machines, M_1, M_2 et M_3, produisent respectivement 1 %, 3 % et 5 % d'articles défectueux. De plus, la machine M_1 produit deux fois plus d'articles dans une journée que la machine M_2, qui elle en produit trois fois plus que la machine M_3. Un article est pris au hasard parmi les articles fabriqués lors d'une journée donnée, puis un deuxième article est pris au hasard parmi ceux fabriqués par la machine qui a produit le premier article sélectionné. Sachant que le premier article examiné est défectueux, quelle est la probabilité que le deuxième article le soit aussi?

Question n^{o} 22

Un appareil est constitué de cinq composants reliés comme dans le diagramme de la figure 2.16. Chaque composant fonctionne avec une probabilité de 0,9, indépendamment des autres composants.

(a) Sachant que le composant n^{o} 1 est en panne, quelle est la probabilité que l'appareil fonctionne?

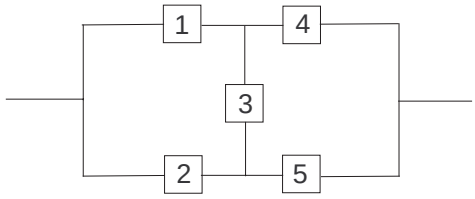

Fig. 2.16. Figure pour l'exercice n° 22

(b) Sachant que le composant n° 1 est en panne et que l'appareil fonctionne malgré tout, quelle est la probabilité que le composant n° 3 fonctionne?

Question n° 23

Avant qu'ils soient déclarés conformes aux spécifications techniques, des appareils doivent passer avec succès deux tests de contrôle de la qualité. D'après les données qui ont été recueillies, 75 % des appareils testés pendant une semaine donnée ont passé avec succès le premier test. Les appareils doivent subir le second test, peu importe qu'ils aient réussi le premier test ou non. On a trouvé que 80 % des appareils qui ont réussi le second test avaient également passé le premier test avec succès. De plus, 20 % de ceux qui ont échoué au second test avaient réussi le premier.

(a) Quelle est la probabilité qu'un appareil donné ait réussi le second test?

(b) Quelle est la probabilité que, pour un appareil donné, le second test contredise le premier?

(c) Quelle est la probabilité qu'un appareil donné ait échoué au second test, sachant qu'il a réussi le premier?

Question n° 24

Dans un certain atelier, la probabilité qu'un article fabriqué satisfasse aux normes en vigueur est égale à 0,9. On propose d'adopter un procédé de contrôle de la qualité qui classifie comme "bons" avec une probabilité de 0,95 les articles effectivement conformes aux normes, et avec une probabilité de 0,15 seulement ceux qui n'y satisfont pas. On décide de faire subir aux articles ce contrôle de la qualité deux fois, de façon indépendante.

(a) Quelle est la probabilité qu'un article ayant subi avec succès à deux reprises le contrôle de la qualité soit effectivement conforme aux normes?

(b) On suppose que, lorsqu'un article ne subit pas avec succès un contrôle, on le retire aussitôt. Soit B_j: un article donné a subi (s'il y a lieu) avec succès le j^{e} contrôle, pour $j = 1, 2$. Calculer (i) $P[B_2]$ et (ii) $P[B_1' \cap B_2']$.

Question n^{o} 25

On dispose de 20 composants de type I, dont 5 sont défectueux, et de 30 composants de type II, dont 15 sont défectueux.

(a) On désire construire un système constitué de 10 composants de type I et de 5 composants de type II placés en série. Quelle est la probabilité que le système fonctionne si les composants sont pris au hasard?

(b) Combien de systèmes différents constitués de quatre composants placés en série, dont au moins deux de type I, peut-on construire si l'on tient compte de l'ordre des composants?

Remarque. On suppose que l'on peut distinguer deux pièces de même type.

Question n^{o} 26

Un système est constitué de n composants, dont les composants A et B.

(a) Montrer que si les composants sont placés en série, alors la probabilité qu'il y ait exactement r composants entre A et B est donnée par

$$\frac{2(n-r-1)}{(n-1)n} \quad \text{pour } r \in \{0,1,\ldots,n-2\}$$

(b) Calculer la probabilité qu'il y ait exactement r composants entre A et B si les composants sont placés en cercle.

(c) Supposons que $n = 5$ et que les composants sont placés en série. Calculer la probabilité que le sous-système constitué des composants A, B et des r composants placés entre eux fonctionne, si les composants fonctionnent indépendamment les uns des autres et ont tous une fiabilité de 0,95.

Question n^{o} 27

Une personne possède une voiture et une motocyclette. La moitié du temps, elle utilise sa moto pour se rendre au travail; le tiers du temps, elle se sert de sa voiture et, le reste du temps, elle utilise les transports en commun. Elle arrive à la maison avant 17 h 30 75 % du temps quand elle utilise sa motocyclette; ce pourcentage est de 60 % quand elle se sert de sa voiture et de 2 % lorsqu'elle utilise les transports en commun. Calculer la probabilité

(a) qu'elle ait utilisé les transports en commun si elle est arrivée à la maison après 17 h 30 lors d'une journée donnée;

(b) qu'elle arrive à la maison avant 17 h 30 lors de deux journées consécutives (indépendantes) et qu'elle utilise les transports en commun au cours d'une seule de ces deux journées.

Question n^{o} 28

Dans un certain aéroport, une navette arrivant du centre-ville s'arrête à chacun des quatre terminaux pour laisser descendre les passagers. On suppose que la probabilité qu'un passager quelconque descende à un terminal donné est de 1/4. S'il y a 20 passagers dans la navette et si ceux-ci occupent des sièges numérotés de 1 à 20, quelle est la probabilité que

(a) les passagers occupant les sièges n^{os} 1 à 4 descendent tous au même arrêt?

(b) les passagers occupant les sièges n^{os} 1 à 4 descendent tous à un arrêt différent?

Question n^{o} 29

Une grille carrée est constituée de 289 points. Une particule se trouve au centre de la grille. Chaque seconde, elle se déplace au hasard sur un des quatre points les plus proches de l'endroit où elle se trouve. Quand la particule arrive à la frontière de la grille, elle est absorbée.

(a) Quelle est la probabilité que la particule soit absorbée au bout de huit secondes?

(b) Soit A_i: la particule se trouve au centre de la grille au bout de i secondes. Calculer $P[A_4]$ (sachant que A_0 est certain, par hypothèse).

Question n^{o} 30

Cinq couples mariés ont acheté 10 billets pour un concert. De combien de façons peuvent-ils s'asseoir si

(a) les cinq hommes veulent être assis ensemble?

(b) les deux conjoints de chaque couple désirent être assis ensemble?

Questions à choix multiple

Question n^{o} 1

Deux semaines avant les dernières élections, un sondage auprès de 1000 électeurs a révélé que 48 % d'entre eux prévoyaient voter pour le parti au pouvoir. Une enquête effectuée après les élections auprès du même échantillon a montré que 90 % des personnes qui ont effectivement voté pour le parti au pouvoir avaient décidé d'avance de voter pour ce parti, tandis que 10 % de celles qui ont voté pour un autre parti avaient pensé (deux semaines avant les élections) voter pour le parti au pouvoir. Soit

A: un électeur, pris au hasard dans l'échantillon, prévoyait voter pour le parti au pouvoir;

B: un électeur, pris au hasard dans l'échantillon, a voté pour le parti au pouvoir.

(A) À partir de l'énoncé du problème, on peut écrire que $P[A] = 0{,}48$ et que
(a) $P[A \cap B] = 0{,}9$; $P[A \cap B'] = 0{,}1$
(b) $P[B \mid A] = 0{,}9$; $P[B' \mid A] = 0{,}1$
(c) $P[A \mid B] = 0{,}9$; $P[A \mid B'] = 0{,}1$
(d) $P[A' \cap B] = 0{,}9$; $P[A \cap B'] = 0{,}1$
(e) $P[A \mid B] = 0{,}9$; $P[B' \mid A] = 0{,}1$

(B) La probabilité de l'événement B est donnée par
(a) 0,45 (b) 0,475 (c) 0,48 (d) 0,485 (e) 0,50 (f) 0,515

(C) Les événements A et B' sont-ils incompatibles?
(a) oui (b) non (c) on ne peut pas conclure à partir de l'information fournie

(D) Les événements A et B' sont-ils indépendants?
(a) oui (b) non (c) on ne peut pas conclure à partir de l'information fournie

(E) Les événements A et B' forment-ils une partition de l'espace échantillon Ω?
(a) oui (b) non (c) on ne peut pas conclure à partir de l'information fournie

(F) Soit E: un électeur, pris au hasard parmi les 1000 qui ont été interrogés, n'a pas voté comme il pensait le faire deux semaines auparavant (en ce qui concerne le parti au pouvoir). On peut écrire que
(a) $P[E] = P[A \mid B'] + P[A' \mid B]$
(b) $P[E] = P[A \cap B] + P[A' \cap B']$
(c) $P[E] = P[B' \mid A] + P[B \mid A']$
(d) $P[E] = P[A \cap B'] + P[A' \cap B']$
(e) $P[E] = P[A \cap B'] + P[A' \cap B]$

Question n° 2

Soit A et B deux événements tels que

$$P[A \cap B] = P\left[A' \cap B\right] = P\left[A \cap B'\right] = p$$

Calculer $P[A \cup B]$.

(a) p (b) $2p$ (c) $3p$ (d) $3p^2$ (e) p^3

Question n° 3

On dispose de neuf composants électroniques, dont un est défectueux. On prend cinq composants au hasard pour construire un système en série. Quelle est la probabilité que le système ne fonctionne pas?

(a) $\frac{1}{3}$ (b) $\frac{4}{9}$ (c) $\frac{1}{2}$ (d) $\frac{5}{9}$ (e) $\frac{2}{3}$

Question n° 4

On lance deux dés simultanément. Si l'on obtient une somme égale à 7 ou 11 avec les dés, alors on lance une pièce de monnaie. Combien y a-t-il de résultats élémentaires (de la forme (dé1, dé2) ou (dé1, dé2, pièce)) dans l'espace échantillon Ω?

(a) 28 (b) 30 (c) 36 (d) 44 (e) 72

Question n° 5

Soit A et B deux événements indépendants tels que $P[A] < P[B]$, $P[A \cap B] = 6/25$ et $P[A \mid B] + P[B \mid A] = 1$. Calculer $P[A]$.

(a) 1/25 (b) 1/5 (c) 6/25 (d) 2/5 (e) 3/5

Question n° 6

Dans une certaine loterie, 4 boules sont tirées sans remise parmi 25 boules numérotées de 1 à 25. On remporte le gros lot si les quatre boules qu'on a choisies sont tirées dans l'ordre qu'on a indiqué. Quelle est la probabilité de remporter le gros lot?

(a) $\frac{1}{12.650}$ (b) $\frac{24}{390.625}$ (c) $\frac{1}{303.600}$ (d) $\frac{1}{390.625}$ (e) $\frac{1}{6.375.600}$

Question n° 7

Les nouvelles plaques d'immatriculation sont constituées de trois lettres suivies de trois chiffres. Si l'on suppose que le I et le O ne sont pas utilisés pour les trois lettres et qu'aucune plaque ne porte les chiffres 000, combien de plaques différentes peut-il y avoir?

(a) $24^3 \times 9^3$ (b) $(26 \times 25 \times 24)(10 \times 9 \times 8)$ (c) $24^3 \times (10 \times 9 \times 8)$
(d) $24^3 \times 999$ (e) $25^3 \times 9^3$

Question n° 8

Soit $P[A \mid B] = 1/2$, $P[B'] = 1/3$ et $P[A \cap B'] = 1/4$. Calculer $P[A]$.

(a) 1/4 (b) 1/3 (c) 5/12 (d) 1/2 (e) 7/12

Question n° 9

Dans la loterie 6/49, 6 boules sont tirées au hasard et sans remise parmi 49 boules numérotées de 1 à 49. Ensuite, une septième boule (le *numéro complémentaire*) est tirée au hasard parmi les 43 boules restantes. Une personne a choisi les six numéros qu'elle croit être les bons pour le tirage. Quelle est la probabilité que cette personne n'ait aucun des sept numéros qui seront tirés (en incluant le complémentaire)?

(a) $\frac{\binom{42}{6}}{\binom{49}{6}}$ (b) $\frac{\binom{42}{7}}{\binom{49}{6}}$ (c) $\frac{\binom{42}{6}}{\binom{49}{7}}$ (d) $\frac{\binom{43}{6}}{\binom{49}{7}}$ (e) $\frac{\binom{42}{7}}{\binom{49}{7}}$

Question n^{o} 10

Lors d'un contrôle de la qualité, on fait subir trois tests à un composant électronique. Après le premier test, le composant est classé selon l'une ou l'autre des catégories suivantes: bon, moyen, ou défectueux. De même, après le deuxième test. Enfin, après le dernier test, le composant est classé soit bon, soit défectueux. Dès qu'un composant est classé défectueux lors d'un test, il est retourné à l'usine pour être réparé. On effectue l'expérience aléatoire suivante: un composant est pris au hasard et l'on observe le résultat du ou des tests qu'il doit subir. Combien y a-t-il de résultats élémentaires dans l'espace échantillon Ω?

(a) 3 (b) 8 (c) 11 (d) 18 (e) 21

Question n^{o} 11

Soit $P[A] = 1/3$, $P[B] = 1/2$, $P[C] = 1/4$, $P[A \mid B] = 1/2$, $P[B \mid A] = 3/4$, $P[A \mid C] = 1/3$, $P[C \mid A] = 1/4$ et $P[B \cap C] = 0$. Calculer la probabilité $P[A \mid B \cup C]$.

(a) 0 (b) 1/3 (c) 4/9 (d) 5/6 (e) 1

Question n^{o} 12

On lance un dé bien équilibré deux fois (de façon indépendante). Soit les événements:

A: les deux nombres obtenus sont différents;

B: le premier nombre obtenu est un "6";

C: les deux nombres obtenus sont pairs.

Quelles paires d'événements sont constituées d'événements indépendants?

(a) aucune paire (b) (A, B) (c) (A, B) et (B, C) (d) (A, B) et (A, C) (e) les trois paires

Question n^{o} 13

On joue plusieurs parties d'un jeu pour lequel la probabilité de gagner une partie quelconque, à partir de la deuxième, est de 3/4 si l'on a remporté la partie précédente et de 1/4 si cela n'est pas le cas. Calculer la probabilité de remporter les parties numéros 2 et 3 de façon consécutive, si la probabilité de remporter la première partie est de 1/2.

(a) 3/16 (b) 1/4 (c) 3/8 (d) 9/16 (e) 5/8

Question n^{o} 14

Un tiroir contient deux pièces de monnaie, dont l'une est non truquée mais dont l'autre possède deux "faces". Une pièce est prise au hasard et on la lance deux fois de façon indépendante. Calculer la probabilité que la pièce non truquée ait été choisie si l'on a obtenu deux fois "face".

(a) 1/5 (b) 1/4 (c) 1/3 (d) 1/2 (e) 3/5

3

Variables aléatoires

Les éléments d'un espace échantillon peuvent prendre diverses formes: nombres réels, mais aussi marque d'un objet, couleur, "bon" ou "défectueux", etc. Comme il est plus facile de travailler avec des nombres réels, dans ce chapitre nous allons transformer tous les résultats élémentaires en valeurs numériques, à l'aide de *variables aléatoires*. Nous allons voir les cas particuliers les plus importants et définir les principales fonctions qui caractérisent les variables aléatoires.

3.1 Introduction

Définition 3.1.1. *Une* **variable aléatoire** *est une fonction à valeurs numériques (réelles) définie sur un espace échantillon.*

Exemple 3.1.1.
(i) Supposons qu'on lance une pièce de monnaie; la fonction X qui associe le nombre 1 au résultat "face" et le nombre 0 au résultat "pile" est une variable aléatoire.

(ii) Supposons maintenant qu'on lance un dé; la fonction X qui associe à chaque résultat élémentaire le nombre obtenu (de sorte que X est la *fonction identité* dans ce cas) est aussi une variable aléatoire. ♢

Exemple 3.1.2. Considérons l'expérience aléatoire qui consiste à observer le temps T qu'une personne doit attendre à un guichet automatique avant de pouvoir s'en servir; la fonction T est une variable aléatoire. ♢

3.1.1 Cas discret

Définition 3.1.2. *Une variable aléatoire est dite de* **type discret** *si le nombre de valeurs différentes qu'elle peut prendre est fini ou infini dénombrable.*

Définition 3.1.3. *La fonction p_X qui associe à chaque valeur possible de la variable aléatoire (discrète) X la probabilité de cette valeur est appelée* **fonction de probabilité** *ou* **masse de probabilité** *de X.*

Soit $\{x_1, x_2, \ldots\}$ l'ensemble des valeurs possibles de la variable aléatoire discrète X; la fonction p_X possède les propriétés suivantes:
(i) $p_X(x_k) \geq 0$ pour tout k;
(ii) $\sum_{k=1}^{\infty} p_X(x_k) = 1$.

Exemple 3.1.1 (suite). (i) Si la pièce de monnaie est bien équilibrée (non truquée), on peut écrire que

x	0	1
$p_X(x)$	1/2	1/2

(ii) De même, si le dé est bien équilibré, alors on a le tableau suivant:

x	1	2	3	4	5	6
$p_X(x)$	1/6	1/6	1/6	1/6	1/6	1/6

♢

Définition 3.1.4. *La fonction F_X qui associe à chaque nombre réel x la probabilité $P[X \leq x]$ que la variable aléatoire X prenne une valeur inférieure ou égale à ce nombre est appelée* **fonction de répartition** *de X. On a:*

$$F_X(x) = \sum_{x_k \leq x} p_X(x_k) \tag{3.1}$$

Remarque. La fonction F_X est *non décroissante* et *continue à droite.*

Exemple 3.1.1 (suite). (i) Dans le cas de la pièce de monnaie, on trouve facilement (si $P[\{\text{face}\}] = 1/2$) que

x	0	1
$F_X(x)$	1/2	1

Remarque. De façon plus complète, on peut écrire que

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1/2 & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

où x est un nombre réel quelconque.
(ii) Si le dé est bien équilibré, alors on déduit de la fonction $p_X(x)$ le tableau suivant:

x	1	2	3	4	5	6
$F_X(x)$	1/6	1/3	1/2	2/3	5/6	1

(voir la figure 3.1). ♢

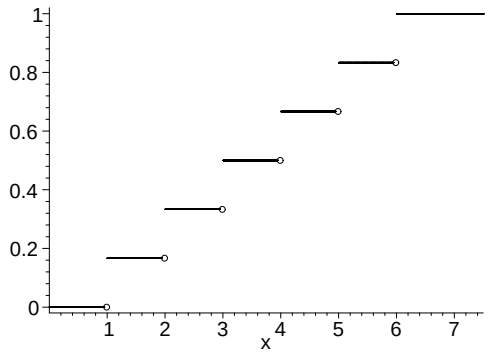

Fig. 3.1. Fonction de répartition de la variable aléatoire dans l'exemple 3.1.1 (ii)

3.1.2 Cas continu

Définition 3.1.5. *Une variable aléatoire qui peut prendre un nombre infini non dénombrable de valeurs est dite variable aléatoire de* **type continu**.

Exemple 3.1.2 (suite). Puisque l'ensemble des valeurs possibles de la variable aléatoire T de l'exemple 3.1.2 est l'intervalle $(0, \infty)$, T est une variable aléatoire *continue*. ♢

Remarque. On suppose dans l'exemple ci-dessus que la personne ne peut pas arriver et commencer immédiatement à se servir du guichet, sinon T serait une variable aléatoire de *type mixte*, qui est une variable à la fois discrète et continue. Nous n'insisterons pas sur ce type de variable aléatoire dans ce manuel.

Définition 3.1.6. *La* **fonction de densité (de probabilité)** *d'une variable aléatoire continue X est une fonction f_X qui possède les propriétés suivantes:*
(i) $f_X(x) \geq 0$ *pour tout nombre réel* x*;*
(ii) $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

La fonction de densité est différente de la fonction de probabilité p_X d'une variable aléatoire discrète. En effet, $f_X(x)$ ne donne *pas* la probabilité que la variable aléatoire X prenne la valeur x. De plus, on peut avoir l'inégalité $f_X(x) > 1$. En fait, on peut écrire que

$$f_X(x)\epsilon \simeq P\left[x - \frac{\epsilon}{2} \leq X \leq x + \frac{\epsilon}{2}\right]$$

où $\epsilon > 0$ est petit. Ainsi, $f_X(x)\,\epsilon$ égale environ la probabilité que X prenne une valeur dans un intervalle de longueur ϵ autour de x.

Définition 3.1.7. *La* **fonction de répartition** F_X *d'une variable aléatoire continue X est définie par*

$$F_X(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(u)\,du \tag{3.2}$$

On déduit de cette définition que

$$\begin{aligned} P[X = x] &= P[x \leq X \leq x] = P[X \leq x] - P[X < x] \\ &= \int_{-\infty}^{x} f_X(u)\,du - \int_{-\infty}^{x^-} f_X(u)\,du = 0 \end{aligned}$$

pour tout nombre réel x, où x^- signifie que le point x n'est pas inclus dans l'intégrale. C'est-à-dire que, *avant* d'effectuer l'expérience aléatoire, la probabilité que l'on obtienne une valeur particulière d'une variable aléatoire *continue* est nulle. Ainsi, si l'on prend un point au hasard dans l'intervalle $[0, 1]$, on peut affirmer que le point que l'on obtiendra n'avait, *a priori*, aucune chance d'être choisi!

On déduit aussi de la définition que

$$\frac{d}{dx} F_X(x) = f_X(x) \tag{3.3}$$

pour tout x où $F_X(x)$ est dérivable.

Remarques.
(i) Si X est une variable aléatoire continue, alors sa fonction de répartition F_X est aussi continue. Cependant, une fonction continue n'est pas nécessairement dérivable en tout point. De plus, la fonction de densité de X peut être une fonction discontinue, comme dans l'exemple suivant. En fait, f_X est une fonction *continue par morceaux*, c'est-à-dire une fonction ayant au plus une nombre fini de *points de saut* (voir la page 2). On dit que f_X possède une point de saut en x_0 si les deux limites $\lim_{x\downarrow x_0} f_X(x)$ et $\lim_{x\uparrow x_0} f_X(x)$ existent, mais sont différentes.

(ii) Toute variable aléatoire X possède une fonction de répartition F_X. Pour simplifier la présentation, nous pourrions théoriquement définir la fonction de densité f_X comme étant la dérivée de F_X, que X soit une variable aléatoire de type discret, continu, ou mixte. Nous avons mentionné dans la remarque précédente que lorsque F_X est une fonction continue, sa dérivée est une fonction continue par morceaux. Cependant, dans le cas d'une variable aléatoire discrète, la fonction de répartition F_X est une *fonction en escalier*, comme dans la figure 3.1. La dérivée d'une fonction en escalier est égale à zéro partout, sauf aux points de saut x_k, $k = 1, 2, \ldots$ On peut écrire que

$$\frac{d}{dx}F_X(x) = \sum_{k=1}^{\infty} P[X = x_k]\delta(x - x_k)$$

où $P[X = x_k] = \lim_{x\downarrow x_k} F_X(x) - \lim_{x\uparrow x_x} F_X(x)$, et $\delta(\cdot)$ est la fonction delta de Dirac (voir la page 5). De façon similaire, la fonction de densité d'une variable aléatoire de type mixte fait intervenir la fonction delta de Dirac. Pour éviter d'utiliser cette *fonction généralisée*, la grande majorité des auteurs préfère considérer les variables (et vecteurs) aléatoires discrètes et continues séparément. Dans le cas discret, on définit la fonction de probabilité $p_X(x_k) = P[X = x_k]$, comme nous l'avons fait ci-dessus, plutôt que la fonction de densité.

Exemple 3.1.3. Supposons que le temps d'attente (en minutes) pour être servi à un guichet dans une banque est une variable aléatoire continue X qui possède la fonction de densité (voir la figure 3.2)

$$f_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1/2 & \text{si } 0 \le x < 1 \\ 3/(2x^4) & \text{si } x \ge 1 \end{cases}$$

Notons que la fonction f_X est bien une fonction de densité, car

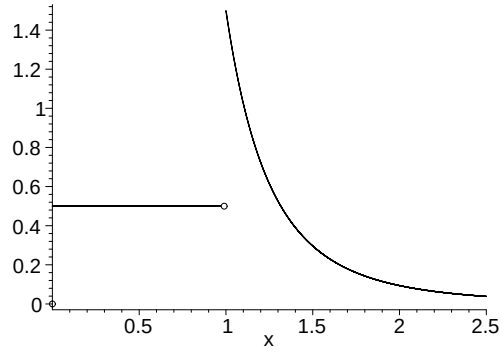

Fig. 3.2. Fonction de densité de la variable aléatoire dans l'exemple 3.1.3

$$\int_0^\infty f_X(x)\,dx = \int_0^1 \frac{1}{2}\,dx + \int_1^\infty \frac{3}{2x^4}\,dx = \frac{1}{2} - \frac{1}{2x^3}\bigg|_1^\infty = \frac{1}{2} + \frac{1}{2} = 1$$

Calculer (a) la fonction de répartition de X et (b) la probabilité conditionnelle $P[X > 2 \mid X > 1]$.

Solution. (a) Par définition,

$$F_X(x) = \begin{cases} \displaystyle\int_{-\infty}^{x} 0\,du = 0 & \text{si } x < 0 \\ \displaystyle\int_{-\infty}^{0} 0\,du + \int_0^x \frac{1}{2}\,du = \frac{x}{2} & \text{si } 0 \le x < 1 \\ \displaystyle\int_{-\infty}^{0} 0\,du + \int_0^1 \frac{1}{2}\,du + \int_1^x \frac{3}{2u^4}\,du = 1 - \frac{1}{2x^3} & \text{si } x \ge 1 \end{cases}$$

(voir la figure 3.3).

(b) On cherche

$$\begin{aligned} P[X > 2 \mid X > 1] &= \frac{P[\{X > 2\} \cap \{X > 1\}]}{P[X > 1]} = \frac{P[X > 2]}{P[X > 1]} \\ &= \frac{1 - P[X \le 2]}{1 - P[X \le 1]} = \frac{1 - F_X(2)}{1 - F_X(1)} = \frac{1 - \frac{15}{16}}{1 - \frac{1}{2}} = \frac{1}{8} \end{aligned}$$

Remarque. Puisque X est une variable aléatoire *continue*, on aurait aussi que $P[X \ge 2 \mid X \ge 1] = \frac{1}{8}$, car $P[X < x] = P[X \le x] = F_X(x)$ pour tout nombre réel x. En général, on a:

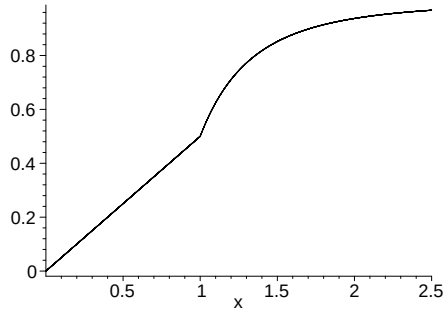

Fig. 3.3. Fonction de répartition de la variable aléatoire dans l'exemple 3.1.3

$$P[a \leq X \leq b] = P[a < X \leq b] = P[a \leq X < b] = P[a < X < b] \tag{3.4}$$

si X est une variable aléatoire *continue*. ♢

3.2 Variables aléatoires discrètes importantes

3.2.1 Distribution binomiale

Supposons que l'on effectue des répétitions d'une expérience aléatoire et que l'on divise l'ensemble des résultats possibles en deux ensembles mutuellement exclusifs et exhaustifs: $\Omega = B_1 \cup B_2$. C'est-à-dire que B_1 et B_2 constituent une partition de l'espace échantillon Ω (chapitre 2, page 44). Si le résultat élémentaire qui se produit appartient à B_1, alors on dit qu'on a obtenu un *succès*; dans le cas contraire, on a obtenu un *échec*.

Définition 3.2.1. *Soit X la variable aléatoire (discrète) qui compte le nombre de succès obtenus au cours de n répétitions d'une expérience aléatoire, où n est fixe. Si*
(i) la probabilité p d'un succès est constante *au cours des n essais et*
(ii) les essais sont indépendants,
alors on dit que X présente (ou suit) une **distribution binomiale** *de* paramètres *n et p. On écrit: $X \sim B(n, p)$.*

Remarque. Un *paramètre* est un symbole qui apparaît dans la définition d'une variable aléatoire et qui peut prendre différentes valeurs. Par exemple, dans le cas de la distribution binomiale, n peut prendre les valeurs $1, 2, \ldots$, et p toutes

les valeurs dans l'intervalle $[0,1]$. En pratique, le paramètre p est généralement inconnu et il faut savoir comment l'*estimer*, comme nous le verrons au chapitre 5 sur la statistique descriptive et l'*estimation*.

Maintenant, supposons que, lors des n essais, on a d'abord obtenu x succès S consécutifs, puis $n-x$ échecs E consécutifs. Par indépendance, on peut écrire que la probabilité de ce résultat élémentaire est

$$P[\underbrace{SS\ldots S}_{x \text{ fois}}\underbrace{EE\ldots E}_{(n-x) \text{ fois}}] = \{P[S]\}^x\{P[E]\}^{n-x} = p^x(1-p)^{n-x}$$

De là, étant donné que l'on peut placer les x succès parmi les n essais de $\binom{n}{x}$ façons différentes, on déduit que la fonction de probabilité de la variable aléatoire $X \sim \mathrm{B}(n,p)$ est donnée par

$$p_X(x) = \binom{n}{x} p^x q^{n-x} \quad \text{pour } x = 0,1,\ldots,n \tag{3.5}$$

où $q := 1-p$.

Remarques.
(i) On a bien: $p_X(x) \geq 0$ pour tout x et, par la *formule du binôme* de Newton,

$$\sum_{x=0}^{n} p_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1$$

(ii) La fonction de répartition de X est

$$F_X(k) = \sum_{x=0}^{k} \binom{n}{x} p^x q^{n-x} \quad \text{pour } k = 0,1,\ldots,n$$

Il n'y a pas de formule simple (sans symbole de sommation) pour F_X. Pour évaluer cette fonction, on peut se servir d'une calculatrice. Des valeurs de la fonction de répartition F_X sont données dans le tableau A.1, en appendice, à la page 514.

(iii) La forme et la position de la fonction de probabilité p_X dépendent des paramètres n et p (voir la figure 3.4).

Exemple 3.2.1. Dans un aéroport, cinq radars sont en service et chaque radar a une probabilité de 0,9 de détecter un avion qui arrive. Les radars fonctionnent indépendamment les uns des autres.

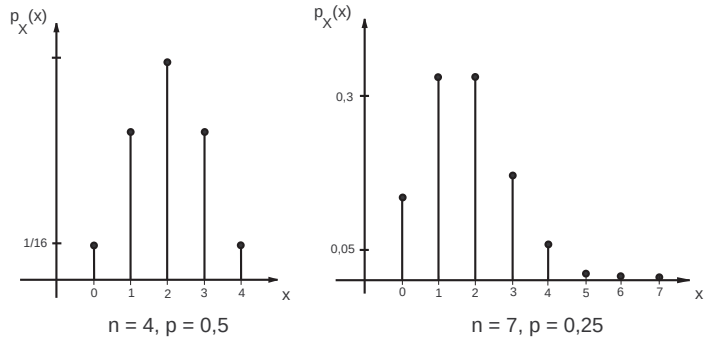

Fig. 3.4. Fonctions de probabilité de variables aléatoires binomiales

(a) Calculer la probabilité qu'un avion qui arrive soit détecté par au moins quatre radars.

(b) Sachant qu'au moins trois radars ont détecté un avion, quelle est la probabilité que les cinq radars aient détecté cet avion?

(c) Combien de radars sont nécessaires, au minimum, si l'on veut que la probabilité qu'un avion qui arrive soit détecté par au moins un radar soit de 0,9999?

Solution. Soit X le nombre de radars qui détectent l'avion.

(a) On a: $X \sim \mathrm{B}(n = 5, p = 0{,}9)$. On cherche

$$P[X \geq 4] = \binom{5}{4}(0{,}9)^4(0{,}1) + (0{,}9)^5 = (0{,}9)^4[5 \times (0{,}1) + 0{,}9] \simeq 0{,}9185$$

Remarque. Soit Y le nombre de radars qui ne détectent *pas* l'avion; on trouve dans le tableau A.1, à la page 514, que

$$P[X \geq 4] = P[Y \leq 1] \simeq 0{,}9185$$

(b) On cherche

$$P[X = 5 \mid X \geq 3] = \frac{P[\{X = 5\} \cap \{X \geq 3\}]}{P[X \geq 3]} = \frac{P[X = 5]}{P[X \geq 3]}$$
$$\overset{\text{(a)}}{=} \frac{P[Y = 0]}{P[Y \leq 2]} \overset{\text{tab. A.1}}{\simeq} \frac{0{,}5905}{0{,}9914} \simeq 0{,}596$$

(c) Maintenant, on a: $X \sim \mathrm{B}(n, p = 0{,}9)$ et on cherche (le plus petit) n tel que $P[X \geq 1] = 0{,}9999$. On a:

$$P[X \geq 1] = 1 - P[X = 0] = 1 - \binom{n}{0}(0{,}9)^0(0{,}1)^{n-0} = 1 - (0{,}1)^n$$
$$= 0{,}9999 \quad \Longleftrightarrow \quad (0{,}1)^n = 0{,}0001 = (0{,}1)^4$$

Donc, on peut écrire que $n_{\min} = 4$.

Remarque. Notons que, étant donné que n est un entier, on ne peut pas, généralement, trouver une valeur de n pour laquelle la probabilité recherchée égale exactement un nombre p donné. Il faut trouver le plus petit n pour lequel la probabilité de l'événement en question est supérieure ou égale à p. Par exemple, ici si l'on avait demandé que la probabilité de détecter l'avion soit de 0,9995, alors la réponse aurait été la même: $n_{\min} = 4$. ♢

3.2.2 Distribution de Bernoulli

Si X présente une distribution binomiale de paramètres $n = 1$ et p, on dit aussi que X présente une **distribution de Bernoulli** de paramètre p. On a donc:

$$\begin{aligned} p_X(x) &= p^x(1-p)^{1-x} \quad \text{si } x = 0, 1 \\ &= \begin{cases} 1-p \text{ si } x = 0 \\ p \quad \text{si } x = 1 \end{cases} \end{aligned} \tag{3.6}$$

On peut aussi écrire que, si les variables aléatoires $X_1, \ldots, X_n$ sont *indépendantes* (voir la section 3.4) et si elles présentent toutes une distribution de Bernoulli de paramètre p, alors

$$X := \sum_{i=1}^{n} X_i \sim \mathrm{B}(n, p)$$

Remarque. On dit que la variable aléatoire binomiale compte le nombre de succès au cours de n *essais de Bernoulli*, c'est-à-dire au cours de n essais indépendants et pour lesquels la probabilité p de succès est la même d'un essai à l'autre.

3.2.3 Distributions géométrique et binomiale négative

Définition 3.2.2. *Soit X la variable aléatoire qui compte le nombre d'essais de Bernoulli effectués jusqu'à ce que l'on obtienne un premier succès. On dit que X présente une* **distribution géométrique** *de paramètre p, où p est la probabilité d'un succès. On écrit: $X \sim Géom(p)$.*

On a:

$$p_X(x) = P[\underbrace{EE\ldots E}_{(x-1)\text{ fois}}S] \overset{\text{ind.}}{=} \{P[E]\}^{x-1}P[S] = q^{x-1}p \tag{3.7}$$

pour $x = 1, 2, \ldots$ On voit que la fonction p_X est décroissante (voir la figure 3.5).

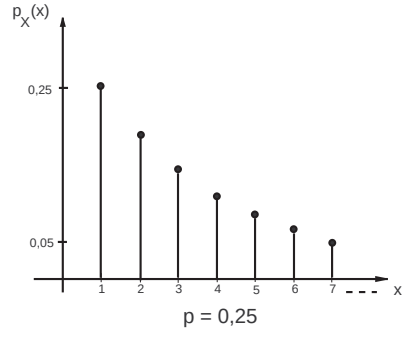

Fig. 3.5. Fonction de probabilité d'une variable aléatoire géométrique

Remarques.
(i) On a: $p_X(x) \geq 0$ pour tout x et

$$\sum_{x=1}^{\infty} p_X(x) = p\sum_{x=1}^{\infty} q^{x-1} = p\frac{1}{1-q} = 1$$

(ii) La fonction de répartition de X est donnée par

$$F_X(x) = \sum_{k=1}^{x} p_X(k) = p\sum_{k=1}^{x} q^{k-1} = p\frac{1-q^x}{1-q} = 1 - q^x \tag{3.8}$$

pour $x = 1, 2, \ldots$ Notons que l'on obtient alors que $P[X > x] = q^x$.
(iii) En utilisant la formule $P[X > x] = q^x$, on montre que

$$P[X > x + y \mid X > x] = P[X > y] \quad \text{pour tous } x, y \in \{1, 2, \ldots\}$$

Cette propriété est appelée *propriété de non-vieillissement* de la distribution géométrique.

Remarque. On définit parfois la distribution géométrique comme étant le nombre d'essais de Bernoulli effectués *avant* d'obtenir un premier succès.

Définition 3.2.3. *Soit X la variable aléatoire qui compte le nombre d'essais de Bernoulli effectués jusqu'à ce que l'on obtienne un r^e succès, où $r = 1, 2, \ldots$ On dit que X présente une* **distribution binomiale négative** *de paramètres r et p. On écrit: $X \sim BN(r, p)$.*

Notons que la distribution géométrique est le cas particulier de la distribution binomiale négative obtenu avec $r = 1$. On obtient la fonction de probabilité de X comme suit:

$$\begin{aligned} p_X(x) &= P[\underbrace{E \ldots E}_{(x-r) \text{ fois}} \underbrace{S \ldots S}_{r \text{ fois}}] + P[\underbrace{E \ldots E}_{(x-r-1) \text{ fois}} SE \underbrace{S \ldots S}_{(r-1) \text{ fois}}] + \\ &\quad \ldots + P[\underbrace{S \ldots S}_{(r-1) \text{ fois}} \underbrace{E \ldots E}_{(x-r) \text{ fois}} S] \\ &= \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad \text{pour } x = r, r+1, \ldots \end{aligned} \tag{3.9}$$

par indépendance et incompatibilité, car il y a $\binom{x-1}{r-1}$ façons de placer les $r-1$ succès au cours des $x-1$ premiers essais (le dernier des x essais devant être un succès).

Remarques.
(i) La distribution binomiale négative est aussi connue sous le nom de distribution de Pascal.

(ii) Comme dans le cas de la distribution binomiale, la forme et la position de la fonction p_X varient selon les valeurs prises par les paramètres r et p.

Exemple 3.2.2. Une personne tire sur une cible jusqu'à ce qu'elle l'ait atteinte deux fois. On suppose que la probabilité qu'un tir atteigne la cible est égale à 0,8. Quelle est la probabilité que la personne doive tirer exactement quatre fois?

Solution. Soit X le nombre de tirs nécessaires pour terminer l'expérience aléatoire; alors, si l'on suppose que les tirs sont indépendants, on peut écrire que $X \sim \text{BN}(r = 2,\ p = 0{,}8)$. On cherche

$$P[X = 4] \equiv p_X(4) = \binom{3}{1} (0{,}8)^2 (1 - 0{,}8)^2 = 3 \times (0{,}64)(0{,}04) = 0{,}0768$$

Remarque. Si la personne cesse de tirer dès qu'elle atteint la cible, alors $X \sim$ Géom$(p = 0{,}8)$ et

$$P[X = 4] = (1 - 0{,}8)^3(0{,}8) = 0{,}0064$$

$\diamondsuit$

3.2.4 Distribution hypergéométrique

Supposons que l'on effectue n répétitions d'une expérience aléatoire mais que la probabilité d'un succès varie d'une répétition à l'autre. Par exemple, on prend *sans* remise n objets dans un lot de taille *finie* N, et on compte le nombre d'objets défectueux obtenus. Dans ce cas, on ne peut pas utiliser la distribution binomiale. Supposons que d objets, parmi les N, sont défectueux (ou bien possèdent une particularité quelconque). Soit X le nombre d'objets défectueux obtenus parmi les n prélevés; on a:

$$p_X(x) = \frac{\binom{d}{x} \cdot \binom{N-d}{n-x}}{\binom{N}{n}} \quad \text{pour } x = 0, 1, \ldots, n \tag{3.10}$$

En effet, il y a $\binom{N}{n}$ *échantillons* différents de taille n et, parmi ceux-ci, il y en a $\binom{d}{x} \cdot \binom{N-d}{n-x}$ qui contiennent exactement x objets défectueux et $n - x$ objets non défectueux.

Remarque. On a: $\binom{n}{k} = 0$ si $k < 0$ ou $k > n$.

Définition 3.2.4. *On dit que la variable aléatoire X, dont la fonction de probabilité est donnée par la formule (3.10), présente une* **distribution hypergéométrique** *de paramètres N, n et d. On écrit: $X \sim Hyp(N, n, d)$.*

On doit avoir: $N \in \{1, 2, \ldots\}$, $n \in \{1, 2, \ldots, N\}$ et $d \in \{0, 1, \ldots, N\}$. De plus, si la taille N du lot est grande comparativement à la taille n de l'échantillon, alors le fait de prendre les objets sans remise n'aura pas beaucoup d'influence sur la probabilité d'obtenir un objet défectueux d'un tirage à l'autre. C'est-à-dire que c'est presque comme si l'on prenait les objets *avec* remise. Or, dans ce cas, on peut écrire que $X \sim \mathrm{B}(n, p = d/N)$. De là, on déduit que l'on peut utiliser la distribution binomiale pour approcher la distribution hypergéométrique. L'approximation obtenue devrait être bonne lorsque $n/N < 0{,}1$.

Remarque. La quantité n/N est appelée *fraction d'échantillonnage.*

Exemple 3.2.3. Des lots de 25 appareils sont soumis au *plan d'échantillonnage* suivant: un échantillon de 5 appareils est prélevé au hasard et sans remise et le lot est accepté si et seulement si l'échantillon contient moins de 3 appareils défectueux. Calculer, en supposant que le lot contient quatre appareils défectueux,
(a) la probabilité que le lot soit accepté;
(b) une approximation de la probabilité calculée en (a) à l'aide d'une distribution binomiale.

Solution. Soit X le nombre d'appareils défectueux dans l'échantillon. On a: $X \sim \text{Hyp}(N = 25, n = 5, d = 4)$.
(a) Soit F: le lot est accepté. On cherche

$$P[F] = P[X \leq 2] = \sum_{x=0}^{2} \frac{\binom{4}{x} \cdot \binom{21}{5-x}}{\binom{25}{5}}$$
$$\simeq 0{,}3830 + 0{,}4506 + 0{,}1502 \simeq 0{,}984$$

(b) On peut écrire que

$$P[X \leq 2] \simeq P[Y \leq 2], \quad \text{où } Y \sim \text{B}(n = 5, p = 4/25)$$
$$= \sum_{y=0}^{2} \binom{5}{y} (4/25)^y (21/25)^{5-y} \simeq 0{,}4182 + 0{,}3983 + 0{,}1517 \simeq 0{,}968$$

Notons qu'ici on a: $n/N = 5/25 = 0{,}2$, ce qui est supérieur à 0,1; donc, on ne s'attendait pas à obtenir une très bonne approximation. ♢

3.2.5 Distribution et processus de Poisson

Soit X une variable aléatoire qui présente une distribution binomiale de paramètres n et p. On peut montrer que, si n tend vers l'infini et p décroît vers 0 de façon telle que le produit np demeure égal à la constante λ, alors la fonction de probabilité de X converge vers la fonction $p_X(x)$ donnée par

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{pour } x = 0, 1, \ldots \tag{3.11}$$

Définition 3.2.5. *On dit qu'une variable aléatoire discrète X dont la fonction de probabilité est donnée par la formule (3.11) présente une* **distribution de Poisson** *de paramètre $\lambda > 0$. On écrit: $X \sim Poi(\lambda)$.*

Remarques.
(i) En utilisant la formule

$$e^x = 1 + x + \frac{x^2}{2!} + \dots$$

on montre facilement que la fonction définie en (3.11) est bien une fonction de probabilité.

(ii) On se sert aussi souvent de la lettre grecque α comme paramètre de la distribution de Poisson. En statistique, on écrit θ pour désigner un paramètre quelconque d'une variable aléatoire.

(iii) L'allure de la fonction de probabilité p_X dépend de la valeur du paramètre λ.

(iv) Pour évaluer la fonction de répartition d'une variable aléatoire de Poisson, on peut se servir d'une calculatrice. Le tableau A.2, à la page 515, donne plusieurs valeurs de cette fonction.

(v) On déduit de ce qui précède que l'on peut utiliser la distribution de Poisson de paramètre $\lambda = np$ pour approcher la distribution binomiale de paramètres n et p. En général, l'**approximation de Poisson** devrait être bonne si $n > 20$ et $p < 0{,}05$. Si la valeur de p est supérieure à $1/2$, alors il faut transformer le nombre de succès en nombre d'échecs avant de faire l'approximation par la distribution de Poisson.

Exemple 3.2.4. Un nouveau type de freins est à l'étude. On pense que ces freins pourraient durer au moins 100.000 km pour 90 % des véhicules qui les utiliseront. Un laboratoire a simulé la conduite de 100 automobiles utilisant ces freins. Soit X le nombre d'automobiles dont les freins ne dureront pas 100.000 km.

(a) Quelle distribution présente X?

(b) On mettra en doute le pourcentage de 90 % si l'on doit changer les freins de 17 automobiles ou plus avant 100.000 km. Quelle est, approximativement, la probabilité d'observer cet événement si, en fait, le pourcentage de 90 % est exact?

Solution. (a) Par définition, si l'on suppose que les automobiles sont indépendantes, alors on peut écrire que $X \sim \mathrm{B}(n = 100, p{=}\ 0{,}10)$.

(b) On cherche $P[X \geq 17]$. On a: $P[X \geq 17] \simeq P[Y \geq 17]$, où $Y \sim \mathrm{Poi}(\lambda = 100(0{,}1){=}10)$. On trouve alors dans le tableau A.2, à la page 515, que

$$P[Y \geq 17] = 1 - P[Y \leq 16] \simeq 1 - 0{,}9730 = 0{,}0270$$

Remarque. Ici, on a: $n = 100$ et $p = 0{,}1$. La valeur de n est très grande, ce qui est préférable, mais celle de p est un peu trop grande pour espérer obtenir une bonne approximation. En fait, on trouve que $P[X \geq 17] \simeq 0{,}021$. ♢

Processus de Poisson

Supposons que la variable aléatoire $N(t)$ désigne le nombre d'événements qui se produiront dans l'intervalle $[0, t]$. Par exemple, on peut s'intéresser au nombre de pannes d'une machine, ou au nombre de clients, ou encore au nombre d'appels téléphoniques dans l'intervalle $[0, t]$. Si l'on fait les hypothèses suivantes:
(i) $N(0) = 0$;
(ii) la valeur de $N(t_4) - N(t_3)$ est indépendante de la valeur prise par $N(t_2) - N(t_1)$ si $0 \leq t_1 < t_2 \leq t_3 < t_4$;
(iii) $N(t + s) - N(t) \sim \text{Poi}(\lambda s)$ pour $s, t \geq 0$,
alors on dit que l'ensemble des variables aléatoires $\{N(t), t \geq 0\}$ est un *processus de Poisson* de taux $\lambda > 0$.

Remarques.
(i) La condition (ii) ci-dessus signifie que tout ce qui se passe dans deux intervalles disjoints est *indépendant*. De plus, la condition (iii) implique que la distribution du nombre d'événements dans un intervalle quelconque ne dépend que de la longueur de cet intervalle. On dit que le processus de Poisson possède des **accroissements indépendants** et des **accroissements stationnaires**, respectivement.
(ii) Les conditions (i) et (iii) impliquent que $N(t) \equiv N(t + 0) - N(0) \sim \text{Poi}(\lambda t)$.
(iii) Le processus de Poisson est un cas particulier très important de ce que l'on appelle *processus stochastique* ou *processus aléatoire*. Ce processus est une *chaîne de Markov à temps continu* particulière. Le processus de Poisson est utilisé à profusion dans la *théorie des communications* et dans la *théorie des files d'attente*.

Exemple 3.2.5. Les appels téléphoniques arrivent à un central selon un processus de Poisson de taux $\lambda = 2$ par minute (c'est-à-dire que les appels arrivent au rythme moyen de deux par minute, selon une distribution de Poisson). Calculer la probabilité qu'exactement deux appels soient reçus pendant chacune des cinq premières minutes d'une heure donnée.

Solution. Soit $N(1)$ le nombre d'appels reçus pendant une période d'une minute. On peut écrire que $N(1) \sim \text{Poi}(2 \cdot 1)$. On calcule d'abord

$$P[N(1) = 2] = P[\text{Poi}(2) = 2] = \frac{e^{-2}2^2}{2!} = 2e^{-2}$$

Ensuite, soit M le nombre de minutes, parmi les cinq minutes considérées, pendant lesquelles exactement deux appels seront reçus. Par indépendance, on peut écrire que M présente une distribution binomiale de paramètres $n = 5$ et $p = 2e^{-2}$. On cherche

$$P[M = 5] = \binom{5}{5}(2e^{-2})^5(1 - 2e^{-2})^{5-5} = 32e^{-10} \simeq 0{,}00145$$

♢

3.3 Variables aléatoires continues importantes

3.3.1 Distribution normale

Définition 3.3.1. *Soit X une variable aléatoire continue qui peut prendre n'importe quelle valeur réelle. Si sa fonction de densité est donnée par*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \textit{pour } -\infty < x < \infty \tag{3.12}$$

alors on dit que X présente une **distribution normale** *(ou* **gaussienne***) de paramètres μ et σ^2, où $\mu \in \mathbb{R}$ et $\sigma > 0$. On écrit: $X \sim N(\mu, \sigma^2)$.*

Remarque. Le paramètre μ est en fait égal à la *moyenne* de X, tandis que σ est l'*écart-type* de X (voir la section 3.5). De plus, l'écart-type d'une variable aléatoire est la racine carrée de la *variance* de cette variable; donc, dans le cas d'une distribution normale, σ^2 est sa variance.

La distribution normale est la distribution continue la plus importante, surtout à cause du *théorème central limite* que l'on verra au chapitre 4. De plus, toutes les distributions normales ont la même forme générale, soit celle d'une *cloche* (voir la figure 3.6).

Les fonctions $f_X(x)$ sont symétriques par rapport à la moyenne μ; c'est-à-dire que

$$f_X(\mu - x) = f_X(\mu + x) \quad \text{pour tout } x \in \mathbb{R}$$

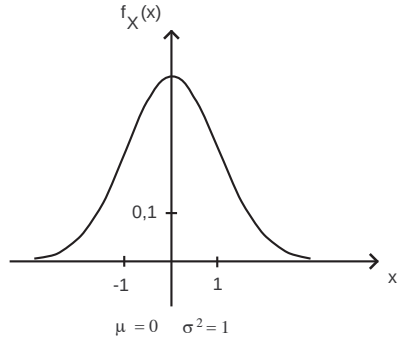

Fig. 3.6. Fonction de densité d'une variable aléatoire normale

Les points $\mu-\sigma$ et $\mu+\sigma$ sont ceux où la fonction f_X change de sens de concavité. Enfin, plus σ est grand, plus la courbe est aplatie; inversement, si σ est petit, alors la courbe est concentrée autour de la moyenne μ.

Maintenant, soit $X \sim \mathrm{N}(\mu, \sigma^2)$; on peut montrer (voir la section 3.4) que si l'on définit $Z = (X - \mu)/\sigma$, alors $Z \sim \mathrm{N}(0, 1)$. La notation Z sera utilisée, en général, pour la distribution normale $\mathrm{N}(0, 1)$. Sa fonction de densité est souvent notée $\phi(z)$.

Remarque. La distribution $\mathrm{N}(0, 1)$ est appelée distribution normale *centrée réduite*. En anglais, elle porte le nom de distribution normale *standard* ou *unit*.

Les principales valeurs de la fonction de répartition de la distribution $\mathrm{N}(0, 1)$, notée $\Phi(z)$, sont présentées dans le tableau A.3, à la page 516. À partir de ce tableau, on peut calculer la probabilité $P[a \leq X \leq b]$ pour n'importe quelle distribution normale. Le tableau donne la valeur de $\Phi(z)$ pour z positif; par symétrie, on peut écrire que $\Phi(-z) = 1 - \Phi(z)$.

Si l'on cherche un nombre a pour lequel $P[X \leq a] = p \geq 1/2$, on trouve d'abord le nombre z dans le tableau A.3 qui correspond à cette probabilité p (il faut parfois interpoler dans le tableau), puis on pose:

$$a = \mu + z \cdot \sigma \tag{3.13}$$

Si $p < 1/2$, la formule devient (par symétrie)

$$a = \mu - z \cdot \sigma \tag{3.14}$$

Finalement, les nombres b qui correspondent aux principales valeurs des probabilités $p := P[X > b]$, par exemple $p = 0{,}05$, $p = 0{,}01$, etc., sont donnés dans

le tableau A.4, à la page 516. Notons que ces nombres peuvent s'écrire comme suit:

$$b = \Phi^{-1}(1-p) \equiv Q^{-1}(p) \tag{3.15}$$

où $Q(z) := 1 - \Phi(z)$ et Q^{-1} est la *fonction inverse* de Q. Les principales valeurs de la fonction $Q^{-1}(p)$ seront aussi utilisées en *statistique*.

Exemple 3.3.1. On suppose que la force de compression X (en livres par pouce carré) d'un certain type de béton présente une distribution normale de paramètres $\mu = 4200$ et $\sigma^2 = (400)^2$.

(a) Calculer la probabilité $P[3000 \leq X \leq 4500]$.

(b) Résoudre les équations suivantes: (i) $P[X \leq a] = 0{,}95$; (ii) $P[X \geq b] = 0{,}90$.

Remarque. En fait, le modèle proposé pour la force de compression du béton ne peut *pas* être le modèle *exact*, car une distribution normale peut prendre n'importe quelle valeur réelle, tandis que la force de compression ne peut pas être négative. Cependant, le modèle en question peut être une bonne approximation du vrai modèle (inconnu). De plus, on trouve que la probabilité qu'une distribution $\mathrm{N}(\mu, \sigma^2)$ prenne une valeur dans l'intervalle $[\mu - 3\sigma, \mu + 3\sigma]$ est supérieure à 99,7 %.

Solution. (a) On a:

$$\begin{aligned} P[3000 \leq X \leq 4500] &= P\left[\frac{3000-4200}{400} \leq \frac{X-4200}{400} \leq \frac{4500-4200}{400}\right] \\ &= P[-3 \leq Z \leq 0{,}75] = \Phi(0{,}75) - \Phi(-3) \\ &\overset{\text{tab. A.3}}{\simeq} 0{,}7734 - 0{,}0013 = 0{,}7721 \end{aligned}$$

(b) (i) On trouve dans le tableau A.3, à la page 516, que $P[Z \leq 1{,}645] \simeq 0{,}95$ (voir aussi le tableau A.4, page 516); alors on peut écrire que

$$a \simeq 4200 + (1{,}645)(400) = 4858$$

(ii) On a: $P[X \geq b] = 0{,}90 \Leftrightarrow P[X < b] = 0{,}1$. Ensuite, on peut écrire que

$$P[X < b] = 0{,}1 \iff P\left[Z < \frac{b-4200}{400}\right] = 0{,}1 \iff Q\left(\frac{b-4200}{400}\right) = 0{,}9$$

Enfin, on trouve dans le tableau A.4 que la valeur qui correspond à $Q^{-1}(0{,}1)$ est environ égale à 1,282. Puisque, par symétrie, $Q^{-1}(0{,}9) = -Q^{-1}(0{,}1)$, il s'ensuit que

$$b \simeq 4200 + (-1{,}282)(400) = 3687{,}2$$

♢

En utilisant le *théorème central limite* (chapitre 4), on peut montrer que, si n est assez grand, on peut se servir d'une distribution normale pour approcher une distribution binomiale de paramètres n et p. Soit $X \sim \mathrm{B}(n, p)$. L'**approximation de Moivre-Laplace** est la suivante:

$$P[X = x] \simeq f_Y(x) \quad \text{pour } x = 0, 1, \ldots, n$$

où $Y \sim \mathrm{N}(np, npq)$.

Remarques.
(i) On remplace donc une *probabilité* en un point par la valeur d'une *fonction de densité* évaluée au même point. Il faut se rappeler que $f_Y(x)$ n'est *pas* la probabilité que Y prenne la valeur x; en effet, puisque Y est une variable aléatoire continue, on sait que $P[Y = x] = 0$ pour tout $x \in \mathbb{R}$.

(ii) Nous verrons à la section suivante que la *moyenne* d'une distribution binomiale est donnée par np, et sa *variance* par $np(1 - p)$. Il est donc logique de choisir $\mu_Y = np$ et $\sigma_Y^2 = npq$.

Lorsqu'on désire évaluer une probabilité comme $P[a \leq X \leq b]$ en se servant d'une approximation par la distribution normale, on utilise la formule suivante:

$$P[a \leq X \leq b] \simeq P\left[Z \leq \frac{b + 0{,}5 - np}{\sqrt{npq}}\right] - P\left[Z \leq \frac{a - 0{,}5 - np}{\sqrt{npq}}\right] \tag{3.16}$$

où a et b sont des entiers.

Remarques.
(i) Pour approcher la probabilité $P[a < X < b]$, on doit écrire que

$$\begin{aligned} P[a < X < b] &= P[a + 1 \leq X \leq b - 1] \\ &\simeq P\left[Z \leq \frac{b - 0{,}5 - np}{\sqrt{npq}}\right] - P\left[Z \leq \frac{a + 0{,}5 - np}{\sqrt{npq}}\right] \end{aligned} \tag{3.17}$$

(ii) Le terme 0,5 dans la formule (3.16) (et (3.17)) est une **correction de continuité** que la plupart des auteurs suggèrent d'utiliser, car on remplace une distribution discrète par une distribution continue.

(iii) L'approximation obtenue devrait être bonne si $np \geq 5$ lorsque $p \leq 0{,}5$, ou $n(1 - p) \geq 5$ lorsque $p \geq 0{,}5$. La distribution normale étant symétrique, il est

plus facile d'approcher la distribution d'une variable aléatoire binomiale pour laquelle $p \simeq 1/2$ que d'approcher la distribution d'une variable binomiale qui possède un paramètre p très petit ou très grand. En fait, dans ce cas, on devrait se servir de l'*approximation de Poisson* que l'on a vue à la section précédente.

Exemple 3.3.2. Un procédé de fabrication produit 10 % d'articles défectueux. On prélève au hasard un échantillon de 200 articles. Soit X le nombre d'articles défectueux dans l'échantillon. Utiliser une distribution normale pour calculer (approximativement) la probabilité $P[X = 20]$.

Solution. On peut supposer que $X \sim \mathrm{B}(n = 200, p = 0{,}10)$; ainsi, on a: $np = 20 > 5$. On pose que

$$\begin{aligned} P[X = 20] &\simeq f_Y(20) \quad \text{où } Y \sim \mathrm{N}(20, 18) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{18}} \exp\left\{-\frac{(20-20)^2}{2 \cdot 18}\right\} \simeq 0{,}0904 \end{aligned}$$

On peut aussi procéder comme suit:

$$\begin{aligned} P[X = 20] &= P[20 \le X \le 20] \\ &\simeq P\left[Z \le \frac{20 + 0{,}5 - 20}{\sqrt{18}}\right] - P\left[Z \le \frac{20 - 0{,}5 - 20}{\sqrt{18}}\right] \\ &\simeq \Phi(0{,}12) - \Phi(-0{,}12) \overset{\text{tab. A.3}}{\simeq} 2(0{,}5478) - 1 = 0{,}0956 \end{aligned}$$

Remarques.
(i) La valeur exacte, obtenue en utilisant la distribution binomiale, est

$$P[X = 20] = \binom{200}{20} (0{,}1)^{20} (0{,}9)^{180} \simeq 0{,}0936$$

(ii) L'approximation de Poisson devrait bien fonctionner dans cet exemple, car n est très grand et p est relativement petit; on trouve que

$$P[X = 20] \simeq P[\mathrm{Poi}(\lambda = 20) = 20] = e^{-20} \frac{20^{20}}{20!} \simeq 0{,}0888$$

Donc, l'*approximation normale* est en fait supérieure dans cet exemple. ♢

3.3.2 Distribution gamma

Définition 3.3.2. *La* **fonction gamma**, *notée* Γ, *est définie par*

$$\Gamma(u) = \int_0^\infty x^{u-1} e^{-x}\, dx \quad \textit{pour } u > 0 \tag{3.18}$$

On peut montrer que $\Gamma(u) = (u-1)\Gamma(u-1)$ si $u > 1$. Puisque l'on trouve directement que $\Gamma(1) = 1$, on peut écrire que

$$\Gamma(u) = (u-1)! \quad \text{si } u \in \{1, 2, \ldots\} \tag{3.19}$$

De plus, on a: $\Gamma(1/2) = \sqrt{\pi}$.

Définition 3.3.3. *Soit* X *une variable aléatoire continue dont la fonction de densité est donnée par*

$$f_X(x) = \frac{\lambda}{\Gamma(\alpha)} (\lambda x)^{\alpha-1} e^{-\lambda x} \quad \textit{pour } x > 0 \tag{3.20}$$

On dit que X *présente une* **distribution gamma** *de paramètres* $\alpha > 0$ *et* $\lambda > 0$. *On écrit:* $X \sim G(\alpha, \lambda)$.

Remarque. Le paramètre λ est un paramètre d'*échelle*, tandis que α est un paramètre de *forme*. La forme de la fonction de densité f_X varie beaucoup avec α, lorsque α est relativement petit (voir la figure 3.7). Lorsque α devient grand, $f_X(x)$ tend vers une densité normale; cela est une conséquence du *théorème central limite*, car, lorsque α est un entier, la variable aléatoire X peut être représentée comme une somme de α variables aléatoires (voir la remarque, page 93).

En général, on ne peut pas donner une formule simple, c'est-à-dire sans signe d'intégrale, pour la fonction de répartition d'une variable aléatoire qui présente une distribution gamma. Cependant, si le paramètre α est un entier (naturel) n, on peut montrer que

$$P[\mathrm{G}(n, \lambda) \le x] = P[\mathrm{Poi}(\lambda x) \ge n] = 1 - P[\mathrm{Poi}(\lambda x) \le n - 1] \tag{3.21}$$

Notons que cette formule nous permet d'exprimer la fonction de répartition d'une variable aléatoire *continue* $X \sim \mathrm{G}(n, \lambda)$ en fonction de celle d'une variable aléatoire *discrète* $Y \sim \mathrm{Poi}(\lambda x)$:

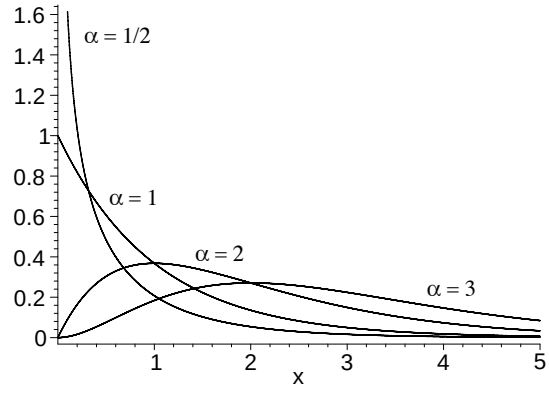

Fig. 3.7. Fonctions de densité de diverses variables aléatoires qui présentent une distribution gamma avec $\lambda = 1$

$$F_X(x) = 1 - F_Y(n-1) \tag{3.22}$$

Cas particuliers

(i) Si α est un entier naturel, alors on dit aussi que X présente une **distribution d'Erlang**, laquelle est importante dans la *théorie des files d'attente*.

(ii) Si $\alpha = n/2$, où $n \in \{1, 2, \ldots\}$, et $\lambda = 1/2$, alors la distribution gamma porte aussi le nom de **distribution du khi-deux** (ou **khi-carré**) à n **degrés de liberté**. On écrit: $X \sim \chi^2_n$. Cette distribution est très utile en *statistique* et nous la reverrons au chapitre 5.

(iii) Si $\alpha = 1$, alors la fonction de densité f_X devient

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{pour } x > 0 \tag{3.23}$$

On dit que X présente une **distribution exponentielle** de paramètre $\lambda > 0$. On écrit: $X \sim \text{Exp}(\lambda)$.

Remarque. Nous verrons au chapitre 4 que, si $X_1, \ldots, X_n$ sont des variables aléatoires *indépendantes* et si $X_i \sim \text{Exp}(\lambda)$ pour $i = 1, \ldots, n$, alors $Y := \sum_{i=1}^{n} X_i$ présente une distribution gamma de paramètres n et λ.

Maintenant, supposons que $\{N(t), t \geq 0\}$ est un processus de Poisson de taux λ. Soit T l'instant d'arrivée du premier événement du processus; on peut écrire que

$$P[T > t] = P[N(t) = 0] = P[\text{Poi}(\lambda t) = 0] = e^{-\lambda t} \tag{3.24}$$

Il s'ensuit que

$$f_T(t) = -\frac{d}{dt}P[T > t] = \lambda e^{-\lambda t} \quad \text{pour } t > 0 \tag{3.25}$$

Ainsi, on peut affirmer que la variable aléatoire T présente une distribution exponentielle de paramètre λ. En utilisant la remarque ci-dessus et les propriétés du processus de Poisson, on peut aussi affirmer, de façon plus générale, que le temps requis pour obtenir n événements (à partir de n'importe quel instant) présente une distribution $\mathrm{G}(n, \lambda)$. Ce résultat nous permet de justifier la formule (3.21).

Remarque. On peut montrer que les distributions exponentielles, comme les distributions géométriques, possèdent la *propriété de non-vieillissement*. C'est-à-dire que si $X \sim \mathrm{Exp}(\lambda)$, alors

$$P[X > t + s \mid X > t] = P[X > s] \quad \text{pour } s, t \geq 0 \tag{3.26}$$

En fait, seules les distributions géométriques et exponentielles possèdent cette propriété de non-vieillissement. De plus, dans le cas de la distribution géométrique, la propriété n'est valide que pour des s et t qui sont des entiers naturels.

Exemple 3.3.3. La durée de vie (en années) d'un appareil de radio présente une distribution exponentielle de paramètre $\lambda = 1/10$. Si l'on achète un appareil vieux de 5 ans, quelle est la probabilité qu'il fonctionne moins de 10 années additionnelles?

Solution. Soit X la durée de vie de l'appareil; on a: $X \sim \mathrm{Exp}(\lambda = 1/10)$. On cherche

$$\begin{aligned} P[X < 15 \mid X > 5] &= 1 - P[X \geq 15 \mid X > 5] = 1 - P[X > 10] = P[X \leq 10] \\ &= \int_0^{10} \frac{1}{10} e^{-x/10}\, dx = -e^{-x/10}\Big|_0^{10} = 1 - e^{-1} \simeq 0{,}6321 \end{aligned}$$

♢

À cause de sa propriété de non-vieillissement, la distribution exponentielle est beaucoup utilisée en *fiabilité*. Cette propriété implique que le *taux de panne* d'un appareil est constant dans le temps. La distribution exponentielle apparaît également dans les *processus stochastiques* et la *théorie des files d'attente*.

Une extension de la distribution exponentielle à toute la droite réelle est obtenue en définissant

$$f_X(x) = \frac{\lambda}{2} e^{-\lambda|x|} \quad \text{pour } -\infty < x < \infty$$

où λ est une constante positive. On dit que la variable aléatoire X présente une **distribution exponentielle double**, ou encore une **distribution de Laplace**, de paramètre λ.

3.3.3 Distribution de Weibull

Définition 3.3.4. *Soit X une variable aléatoire continue dont la fonction de densité est de la forme*

$$f_X(x) = \lambda\beta x^{\beta-1} \exp\left(-\lambda x^{\beta}\right) \quad \textit{pour } x > 0 \tag{3.27}$$

On dit que X présente une **distribution de Weibull** *de paramètres $\lambda > 0$ et $\beta > 0$. On écrit: $X \sim W(\lambda, \beta)$.*

La distribution de Weibull généralise la distribution exponentielle, qui est obtenue en prenant $\beta = 1$. Elle est importante en fiabilité. Comme la distribution gamma, elle peut être utilisée dans de nombreuses applications à cause des diverses formes que prend sa fonction de densité selon les valeurs données au paramètre β. C'est aussi une des distributions qui portent le nom de *distributions extrêmes*. Ces distributions sont utilisées dans la modélisation de phénomènes qui se produisent très peu souvent, comme des températures extrêmement froides ou chaudes, des crues de rivières exceptionnelles, etc.

3.3.4 Distribution bêta

Définition 3.3.5. *Soit X une variable aléatoire continue dont la fonction de densité est donnée par*

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \textit{pour } 0 < x < 1 \tag{3.28}$$

où $\alpha > 0$ et $\beta > 0$. On dit que X présente une **distribution bêta** *de paramètres α et β. On écrit: $X \sim Be(\alpha, \beta)$.*

Si $X \sim \mathrm{Be}(\alpha, \beta)$ et $Y := a + (b-a)X$, où $a < b$, on dit que Y présente une distribution *bêta généralisée*.

Cas particulier

Soit $X \sim \mathrm{Be}(\alpha, \beta)$. Si $\alpha = \beta = 1$, alors on a:

$$f_X(x) = 1 \quad \text{pour } 0 < x < 1 \tag{3.29}$$

On dit que la variable aléatoire continue X présente une **distribution uniforme** sur l'intervalle $(0, 1)$. On écrit: $X \sim \mathrm{U}(0, 1)$. En général, on a: $X \sim \mathrm{U}(a, b)$ si (voir la figure 3.8)

$$f_X(x) = \frac{1}{b-a} \quad \text{pour } a < x < b \tag{3.30}$$

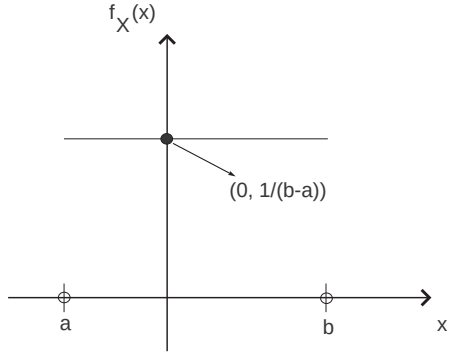

Fig. 3.8. Fonction de densité de probabilité d'une variable aléatoire uniforme sur l'intervalle (a, b)

Remarques.
(i) Cette fonction de densité est obtenue, par exemple, lorsqu'on prend un point *au hasard* dans l'intervalle (a, b). Comme la probabilité que le point choisi soit proche de x, où $a < x < b$, est la même pour tout x, la fonction $f_X(x)$ doit être *constante* dans l'intervalle (a, b). Notons qu'une variable aléatoire qui présente une distribution uniforme sur l'intervalle (a, b) présente aussi une distribution bêta généralisée de paramètres $\alpha = \beta = 1$.

(ii) On peut montrer que, s'il y a eu exactement un événement d'un processus de Poisson dans l'intervalle $[0, t]$, alors l'instant T où cet événement s'est produit présente une distribution uniforme sur $[0, t]$. C'est-à-dire que

$$T \mid \{N(t) = 1\} \sim \mathrm{U}[0, t] \tag{3.31}$$

Exemple 3.3.4. Un point est pris au hasard sur un segment de longueur L. Quelle est la probabilité que la longueur du petit segment divisée par la longueur du grand segment soit inférieure à 1/4?

Solution. Supposons, sans perte de généralité, que le segment commence à 0. Soit X le point choisi. Alors $X \sim \mathrm{U}[0, L]$.
(i) Si $X \in [0, L/2]$, alors on doit avoir:

$$\frac{X}{L-X} < \frac{1}{4} \quad \Longleftrightarrow \quad 4X < L - X \quad \Longleftrightarrow \quad X < \frac{L}{5}$$

(ii) Si $X \in (L/2, L]$, alors il faut que

$$\frac{L-X}{X} < \frac{1}{4} \quad \Longleftrightarrow \quad 4L - 4X < X \quad \Longleftrightarrow \quad X > \frac{4L}{5}$$

On a:

$$P[X < L/5] = \int_0^{L/5} \frac{1}{L-0}\, dx = \frac{L/5}{L} = \frac{1}{5}$$

De même,

$$P[X > 4L/5] = \int_{4L/5}^{L} \frac{1}{L}\, dx = \frac{L - (4L/5)}{L} = \frac{1}{5}$$

Donc, la probabilité recherchée égale $\frac{1}{5} + \frac{1}{5} = \frac{2}{5}$.
Remarques.
(i) Par symétrie, il suffisait de considérer un seul des deux cas et de multiplier la probabilité obtenue par 2.
(ii) Comme on le voit dans cet exemple, la probabilité que la variable aléatoire uniforme X soit située dans un sous-intervalle ne dépend que de la longueur de ce sous-intervalle. ♢

3.3.5 Distribution lognormale

Définition 3.3.6. *Soit X une variable aléatoire continue qui ne prend que des valeurs positives. Si $Y := \ln X$ présente une distribution $N(\mu, \sigma^2)$, alors on dit que X présente une* **distribution lognormale** *de paramètres μ et σ^2. On écrit: $X \sim LN(\mu, \sigma^2)$. La fonction de densité de X est donnée par*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{ -\frac{(\ln x - \mu)^2}{2\sigma^2} \right\} \quad \textit{pour } x > 0 \tag{3.32}$$

où $\mu \in \mathbb{R}$ et $\sigma > 0$.

Remarque. Dans plusieurs situations, la distribution lognormale peut être un modèle plus réaliste que la distribution normale, car elle est toujours positive. Par exemple, le poids d'objets manufacturés pourrait présenter une distribution lognormale.

Exemple 3.3.5. Soit $X \sim \mathrm{LN}(5,4)$. Calculer $P[X \leq 100]$.

Solution. On a:

$$\begin{aligned} P[X \leq 100] &= P[\ln X \leq \ln 100] = P[Y \leq \ln 100], \quad \text{où } Y \sim \mathrm{N}(5,4) \\ &= P\left[Z \leq \frac{\ln 100 - 5}{2}\right] \simeq \Phi(-0{,}2) \overset{\text{tab. A.3}}{\simeq} 0{,}4207 \end{aligned}$$

♢

3.4 Fonctions de variables aléatoires

Étant donné qu'une variable aléatoire est une *fonction* à valeurs réelles et que la *composition* de deux fonctions est une autre fonction, on peut affirmer que si X est une variable aléatoire, alors $Y := g(X)$, où g est une fonction à valeurs réelles définie sur la droite réelle, est également une variable aléatoire. Dans cette section, nous allons voir comment obtenir la fonction de probabilité ou la fonction de densité de Y.

3.4.1 Cas discret

Puisqu'une *fonction* g associe un seul nombre réel à chaque valeur possible de la variable aléatoire X, on peut affirmer que si X est une variable de type discret, alors $Y = g(X)$ sera aussi une variable aléatoire discrète, quelle que soit la fonction g. En effet, Y ne peut pas prendre plus de valeurs différentes que X. Pour obtenir la fonction de probabilité de Y, on applique la transformation g à chacune des valeurs possibles de X et on additionne les probabilités des valeurs x de la variable aléatoire X qui correspondent au même y.

Exemple 3.4.1. Soit X une variable aléatoire discrète dont la fonction de probabilité est donnée par

x	-1	0	1
$p_X(x)$	$1/4$	$1/4$	$1/2$

On définit $Y = 2X$. Puisque la fonction $g : x \to 2x$ est *bijective* (c'est-à-dire qu'à un $y = g(x) = 2x$ correspond un et un seul x et vice versa), le nombre de valeurs possibles de la variable aléatoire Y sera le même que le nombre de valeurs possibles de X. On trouve que

y	-2	0	2	Σ
$p_Y(y)$	1/4	1/4	1/2	1

Soit maintenant $W = X^2$. Étant donné qu'à deux valeurs de X, soit -1 et 1, correspond la même valeur $w = 1$, il faut additionner $p_X(-1)$ et $p_X(1)$ pour obtenir $p_W(1)$. On a donc:

w	0	1	Σ
$p_W(w)$	1/4	3/4	1

♢

Dans le cas où la variable aléatoire X peut prendre un nombre infini (dénombrable) de valeurs, on applique la transformation à chaque valeur possible de X, et on essaie de trouver une formule générale pour la fonction $p_Y(y)$.

Exemple 3.4.2. Soit $X \sim \text{Géom}(p)$ et $Y = X^2$. Puisque $X \in \{1, 2, \ldots\}$, la transformation est bijective et l'on calcule facilement

$$p_Y(y) = p_X(\sqrt{y}) = q^{\sqrt{y}-1}p \quad \text{pour } y = 1, 4, 9, \ldots$$

♢

3.4.2 Cas continu

La *composition* de deux fonctions continues est une autre fonction continue. Par conséquent, si X est une variable aléatoire continue et g est une fonction continue, alors $Y := g(X)$ est également une variable aléatoire continue. Dans le cadre de ce livre, nous n'allons considérer que le cas où la fonction g est *bijective*. Dans ce cas, la fonction inverse $g^{-1}(y)$ existe et on peut se servir de la proposition suivante pour obtenir la fonction de densité de la nouvelle variable aléatoire Y.

Proposition 3.4.1. *Supposons que l'équation $y = g(x)$ possède une solution (réelle) unique: $x = g^{-1}(y)$. Alors la fonction de densité de Y est donnée par la formule*

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \tag{3.33}$$

Exemple 3.4.3. On peut utiliser la proposition précédente pour démontrer le résultat que nous avons énoncé à la section 3.3: si X présente une distribution $\mathrm{N}(\mu, \sigma^2)$, alors $Z := (X-\mu)/\sigma$ présente une distribution normale centrée réduite. En effet, on a:

$$z = g(x) = (x - \mu)/\sigma \Leftrightarrow g^{-1}(z) = \mu + \sigma z$$

Il s'ensuit que

$$f_Z(z) = f_X(\mu + \sigma z) \left| \frac{d}{dz}(\mu + \sigma z) \right| = \frac{1}{\sqrt{2\pi}\sigma} \exp(-z^2/2)|\sigma| = \phi(z)$$

pour $-\infty < z < \infty$, car $\sigma > 0$. De même, on montre que si $Y := aX + b$, alors Y présente une distribution normale de paramètres $a\mu + b$ et $a^2\sigma^2$. ♢

Exemple 3.4.4. Soit $X \sim \mathrm{U}(0, 1)$ et $Y := -\theta \ln X$, où $\theta > 0$. Notons d'abord que les valeurs possibles de la variable aléatoire Y sont celles situées dans l'intervalle $(0, \infty)$. Ensuite, on a: $g^{-1}(y) = e^{-y/\theta}$, de sorte que

$$f_Y(y) = f_X(e^{-y/\theta}) \left| \frac{d}{dy} e^{-y/\theta} \right| = 1 \cdot \left| -\frac{1}{\theta} e^{-y/\theta} \right| = \frac{1}{\theta} e^{-y/\theta}$$

pour $0 < y < \infty$. Ainsi, on peut affirmer que Y présente une distribution exponentielle de paramètre $1/\theta$. On se sert de ce résultat en *simulation*. En effet, si l'on désire générer une observation d'une variable aléatoire qui présente une distribution *exponentielle* de paramètre $\lambda = 2$ (par exemple), il suffit de générer une observation x d'une distribution $\mathrm{U}(0, 1)$, puis d'effectuer la transformation $y = -(1/2) \ln x$. Il n'est donc pas nécessaire d'écrire un programme spécialement pour générer des observations de distributions exponentielles. Les langages informatiques en général, et même beaucoup de calculatrices, permettent de générer des observations pseudo-aléatoires de distributions uniformes. ♢

3.5 Caractéristiques des variables aléatoires

Dans cette section, nous allons présenter quelques quantités numériques qui permettent de caractériser une variable aléatoire X. Toutes ces quantités sont obtenues en calculant l'*espérance mathématique* de diverses fonctions de X.

Définition 3.5.1. *On définit l'***espérance mathématique** *(ou simplement la* **moyenne***) d'une fonction $g(X)$ d'une variable aléatoire X par*

$$E[g(X)] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} g(x_i)p_X(x_i) & \textit{si } X \textit{ est discrète} \\ \displaystyle\int_{-\infty}^{\infty} g(x)f_X(x)\,dx & \textit{si } X \textit{ est continue} \end{cases} \tag{3.34}$$

Propriétés.
(i) $E[c] = c$ pour toute constante réelle c.
(ii) $E[c_1 g(X) + c_0] = c_1 E[g(X)] + c_0$ pour toutes constantes réelles c_1 et c_0.

Remarques.
(i) E est donc un *opérateur linéaire*.
(ii) L'espérance mathématique peut être infinie ou même ne pas exister.
(iii) Si X est une variable aléatoire de *type mixte*, c'est-à-dire une variable aléatoire à la fois discrète et continue, alors on peut calculer $E[g(X)]$ en décomposant le problème en deux parties. Par exemple, supposons qu'on lance une pièce de monnaie avec laquelle la probabilité d'obtenir "pile" est de 3/8. Si l'on obtient "pile", alors la variable aléatoire X prend la valeur 1; sinon, X est un nombre pris au hasard dans l'intervalle $[2, 4]$. On peut obtenir $E[g(X)]$ comme suit:

$$E[g(X)] = 1 \times \frac{3}{8} + \left(\int_2^4 g(x) \cdot \tfrac{1}{2}\,dx \right) \times \frac{5}{8}$$

Définition 3.5.2. *La* **moyenne** *d'une variable aléatoire X est donnée par*

$$\mu_X \equiv E[X] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} x_i p_X(x_i) & \textit{si } X \textit{ est discrète} \\ \displaystyle\int_{-\infty}^{\infty} x f_X(x)\,dx & \textit{si } X \textit{ est continue} \end{cases} \tag{3.35}$$

Exemple 3.5.1. Soit $X \sim \text{Poi}(\lambda)$. On a:

$$\begin{aligned}\mu_X &= \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} = e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= e^{-\lambda}\lambda \left\{ 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \ldots \right\} = e^{-\lambda}\lambda e^{\lambda} = \lambda\end{aligned}$$

♢

Exemple 3.5.2. Soit $X \sim \text{Exp}(\lambda)$. On calcule

$$\mu_X = \int_0^{\infty} x\lambda e^{-\lambda x}\, dx = \lambda \cdot \frac{1}{\lambda^2} = \frac{1}{\lambda}$$

car, en général, si $a > 0$ on a:

$$\begin{aligned}\int_0^{\infty} x^n e^{-ax}\, dx &\overset{y=ax}{=} \int_0^{\infty} \left(\frac{y}{a}\right)^n e^{-y}\, \frac{dy}{a} \\ &= \frac{1}{a^{n+1}} \int_0^{\infty} y^n e^{-y}\, dy = \frac{\Gamma(n+1)}{a^{n+1}} = \frac{n!}{a^{n+1}}\end{aligned} \tag{3.36}$$

Remarque. On pouvait aussi effectuer l'intégrale par parties. ♢

Définition 3.5.3. *Une* **médiane** *d'une variable aléatoire X est une valeur x_m ($\equiv \tilde{x}$) pour laquelle*

$$P[X \leq x_m] \geq \frac{1}{2} \quad \textbf{et} \quad P[X \geq x_m] \geq \frac{1}{2} \tag{3.37}$$

Remarques.
(i) Lorsque X est une variable aléatoire discrète, la médiane n'est pas nécessairement unique. Elle n'est *pas* unique s'il existe un nombre réel a tel que $F_X(a) = 1/2$. Par exemple, soit $X \sim \text{B}(n = 2, p = 1/2)$. On a:

$$F_X(1) = P[X = 0] + P[X = 1] = 1/4 + 1/2 = 3/4$$

Dans ce cas, le nombre $x_m = 1$ satisfait à la définition ci-dessus, puisque

$$P[X \leq 1] = 3/4 \geq 1/2 \quad \text{et} \quad P[X \geq 1] = 3/4 \geq 1/2$$

aussi. De plus, $x_m = 1$ est le seul nombre pour lequel les deux inégalités sont vérifiées à la fois. Par contre, soit

x	1	2	3
$p_X(x)$	1/4	1/4	1/2

Le nombre 2 satisfait à la définition de la médiane, mais le nombre 2,5 (par exemple) également. En fait, tout nombre dans l'intervalle $[2, 3]$ peut être considéré comme la médiane de X. Remarquons que, si l'on change les probabilités comme suit: $P[X = 1] = 1/4$, $P[X = 2] = 1/8$ et $P[X = 3] = 5/8$, alors la médiane $x_m = 3$ est unique. De même, si $P[X = 1] = 1/4$, $P[X = 2] = 3/8$ et $P[X = 3] = 3/8$, alors $x_m = 2$ est l'unique médiane de X.

(ii) Lorsque X est une variable aléatoire continue qui prend toutes ses valeurs dans un seul intervalle (fini ou infini), la médiane *est* unique et on peut la définir comme suit:

$$P[X \leq x_m] = 1/2 \quad (\Longrightarrow \ P[X \geq x_m] = 1/2) \tag{3.38}$$

Exemple 3.5.1 (suite). Supposons $X \sim \text{Poi}(2)$. On trouve dans le tableau A.2, à la page 515, que $P[X \leq 1] \simeq 0{,}4060$ et $P[X \leq 2] \simeq 0{,}6767$. Il n'y a donc aucun nombre x_m tel que $P[X \leq x_m] = P[X \geq x_m] = 1/2$. Cependant, on a:

$$P[X \leq 2] \simeq 0{,}6767 \geq 1/2 \quad \text{et} \quad P[X \geq 2] \simeq 1 - 0{,}4060 \geq 1/2$$

De plus, 2 est le seul nombre pour lequel les *deux* inégalités sont vérifiées. Alors $x_m = 2$ est *la* médiane de X. ♢

Exemple 3.5.2 (suite). Si $X \sim \text{Exp}(\lambda)$, alors on a:

$$P[X \leq x_m] = \int_0^{x_m} \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_0^{x_m} = 1 - e^{-\lambda x_m}$$

Il s'ensuit que

$$P[X \leq x_m] = 1/2 \Leftrightarrow 1 - e^{-\lambda x_m} = 1/2 \Leftrightarrow x_m = \frac{\ln 2}{\lambda}$$

On peut vérifier que l'on a bien:

$$P\left[X \geq \frac{\ln 2}{\lambda}\right] = \int_{\frac{\ln 2}{\lambda}}^{\infty} \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_{\frac{\ln 2}{\lambda}}^{\infty} = e^{-\lambda \frac{\ln 2}{\lambda}} = 1/2$$

Donc, *la* médiane est donnée par $x_m = (\ln 2)/\lambda$. ♢

La médiane est utile lorsque la variable aléatoire X prend des valeurs très grandes (en valeur absolue) par rapport aux autres. En effet, la médiane est moins influencée par ces valeurs extrêmes que la moyenne μ_X.

La médiane, dans le cas continu, est un cas particulier de la notion de *quantile*.

Définition 3.5.4. *Soit X une variable aléatoire continue dont l'ensemble des valeurs possibles est un intervalle quelconque (a, b). Le nombre x_p est appelé $100(1-p)^e$* **quantile** *de X si*

$$P[X \leq x_p] = 1 - p \quad \text{où } 0 < p < 1 \tag{3.39}$$

Si $100p$ est un entier, alors x_p est aussi appelé $100(1-p)^{\text{e}}$ **centile** de X. La médiane d'une variable aléatoire continue est donc le 50^{e} centile de X. Le 25^{e} centile est aussi appelé *premier quartile*, le 50^{e} centile est le *second quartile*, et le 75^{e} centile est le *troisième quartile*. Finalement, la différence entre le troisième et le premier quartile est appelée *écart interquartile*.

Définition 3.5.5. *Un* **mode** *d'une variable aléatoire X est n'importe quelle valeur x qui correspond à un maximum local pour $p_X(x)$ ou $f_X(x)$.*

Remarque. Le mode n'est donc pas nécessairement unique. Une distribution qui possède un seul mode est dite *unimodale*.

Exemple 3.5.1 (suite). Soit $X \sim \text{Poi}(2)$. À partir du tableau A.2, page 515, on obtient que $P[X = 0] \simeq 0{,}135$, $P[X = 1] \simeq 0{,}271$, $P[X = 2] \simeq 0{,}271$, $P[X = 3] \simeq 0{,}180$, etc. Alors X possède deux modes: à $x = 1$ et $x = 2$.

Remarque. On a bien:

$$P[X = 1] = e^{-2}\frac{2^1}{1!} = e^{-2}\frac{2^2}{2!} = P[X = 2] \simeq 0{,}271$$

♢

Exemple 3.5.2 (suite). Soit $X \sim \text{Exp}(\lambda)$. Alors, X étant une variable aléatoire continue, on peut utiliser le calcul différentiel pour calculer son ou ses modes.

On a:

$$\frac{d}{dx} f_X(x) = \frac{d}{dx} \lambda e^{-\lambda x} = -\lambda^2 e^{-\lambda x} \neq 0$$

pour tout $x \in (0, \infty)$. Cependant, puisque $f_X(x)$ est une fonction décroissante, on peut affirmer que le mode de X est à $x = 0^+$ ($x \to \infty$ correspond à un minimum). ♢

Les différentes quantités définies ci-dessus sont des *mesures de position centrale*. Nous allons poursuivre en définissant des *mesures de dispersion*.

Définition 3.5.6. *L'***étendue** *d'une variable aléatoire est la différence entre la plus grande valeur que cette variable peut prendre et la plus petite.*

Par exemple, l'étendue d'une variable aléatoire $X \sim \mathrm{B}(n, p)$ est égale à $n-0 = n$. De même, si $X \sim \mathrm{Exp}(\lambda)$, alors son étendue est égale à $\infty - 0 = \infty$.

Définition 3.5.7. *La* **variance** *d'une variable aléatoire X est définie par*

$$\sigma_X^2 \equiv \mathrm{VAR}[X] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} (x_i - \mu_X)^2 p_X(x_i) & \textit{si } X \textit{ est discrète} \\ \displaystyle\int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx & \textit{si } X \textit{ est continue} \end{cases} \tag{3.40}$$

Remarques.
(i) On déduit immédiatement de la définition que la variance de X est toujours non négative (elle peut être infinie). En fait, elle est strictement positive, sauf si la variable aléatoire X est une constante, qui est une variable aléatoire *dégénérée*. Enfin, plus la variance est grande, plus la variable aléatoire possède une distribution qui est dispersée autour de sa moyenne.

(ii) On définit aussi l'*écart-type* (*standard deviation*, en anglais) d'une variable aléatoire X par

$$\mathrm{STD}[X] = \sqrt{\mathrm{VAR}[X]} \equiv \sigma_X \tag{3.41}$$

On préfère souvent travailler avec l'écart-type plutôt qu'avec la variance d'une variable aléatoire, car il est plus facile à interpréter; en effet, l'écart-type est exprimé dans les mêmes unités de mesure que X, tandis que les unités de la variance sont les unités de X élevées au carré.

Maintenant, on peut écrire que

$$\mathrm{VAR}[X] \equiv E[(X - E[X])^2] \tag{3.42}$$

et on peut montrer que

$$\mathrm{VAR}[X] = E[X^2] - (E[X])^2 \tag{3.43}$$

Exemple 3.5.1 (suite). Lorsque $X \sim \mathrm{Poi}(\lambda)$, on calcule

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

Alors, en utilisant la formule (3.43) avec $E[X] = \lambda$, on trouve que

$$\mathrm{VAR}[X] = (\lambda + \lambda^2) - (\lambda)^2 = \lambda$$

Ainsi, dans le cas de la distribution de Poisson, son paramètre λ est à la fois sa moyenne et sa variance. ♢

Exemple 3.5.2 (suite). On a déjà trouvé que si $X \sim \mathrm{Exp}(\lambda)$, alors $E[X] = 1/\lambda$. On calcule maintenant (voir (3.36))

$$E[X^2] = \int_0^{\infty} x^2 \lambda e^{-\lambda x}\, dx = \lambda \frac{2!}{\lambda^{2+1}} = \frac{2}{\lambda^2}$$

Il s'ensuit que

$$\mathrm{VAR}[X] = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Notons que, dans le cas de la distribution exponentielle, sa moyenne et son écart-type sont égaux. ♢

Le tableau 3.1, à la page 107, donne la moyenne et la variance des différentes distributions de probabilité des sections 3.2 et 3.3.

Tableau 3.1. Moyennes et variances des distributions de probabilité des sections 3.2 et 3.3

Distribution	**Paramètres**	**Moyenne**	**Variance**
Bernoulli	p	p	pq
Binomiale	n et p	np	npq
Hypergéométrique	N, n et d	$n \cdot \frac{d}{N}$	$n \cdot \frac{d}{N} \cdot \left(1 - \frac{d}{N}\right) \cdot \left(\frac{N-n}{N-1}\right)$
Géométrique	p	$\frac{1}{p}$	$\frac{q}{p^2}$
Pascal	r et p	$\frac{r}{p}$	$\frac{rq}{p^2}$
Poisson	λ	λ	λ
Uniforme	$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponentielle	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Laplace	λ	0	$\frac{2}{\lambda^2}$
Gamma	α et λ	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
Weibull	λ et β	$\frac{\Gamma(1+\beta^{-1})}{\lambda^{1/\beta}}$	$\frac{\Gamma(1+2\beta^{-1}) - \Gamma^2(1+\beta^{-1})}{\lambda^{2/\beta}}$
Normale	μ et σ^2	μ	σ^2
Bêta	α et β	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$
Lognormale	μ et σ^2	$e^{\mu+\frac{1}{2}\sigma^2}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

Les propriétés de l'*opérateur mathématique* VAR sont les suivantes:
(i) $\mathrm{VAR}[c] = 0$ pour toute constante c.
(ii) $\mathrm{VAR}[c_1 g(X) + c_0] = c_1^2\ \mathrm{VAR}[g(X)]$ pour toutes constantes réelles c_1 et c_0. Donc, l'opérateur VAR n'est *pas* linéaire.

La moyenne et la variance d'une variable aléatoire sont des cas particuliers de ce que l'on appelle *moments* de cette variable.

Définition 3.5.8. *Le* **moment d'ordre** k *(ou* k^e **moment***) d'une variable aléatoire* X *par rapport à un point* **a** *est défini par*

$$E[(X-a)^k] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} (x_i - a)^k p_X(x_i) & \textit{si } X \textit{ est discrète} \\ \displaystyle\int_{-\infty}^{\infty} (x-a)^k f_X(x)\,dx & \textit{si } X \textit{ est continue} \end{cases} \tag{3.44}$$

Cas particuliers
(i) Les **moments par rapport à l'origine**, ou **moments non centrés**, de X sont:

$$E[X^k] \equiv \mu'_k = \sum_{i=1}^{\infty} x_i^k p_X(x_i) \quad \text{ou} \quad \int_{-\infty}^{\infty} x^k f_X(x)\,dx \tag{3.45}$$

pour $k = 0, 1, \ldots$ On a: $\mu'_0 = 1$ et $\mu'_1 = \mu_X = E[X]$.
(ii) Les **moments par rapport à la moyenne**, ou **moments centrés**, de X sont:

$$E[(X-\mu_X)^k] \equiv \mu_k = \sum_{i=1}^{\infty} (x_i - \mu_X)^k p_X(x_i) \quad \text{ou} \quad \int_{-\infty}^{\infty} (x-\mu_X)^k f_X(x)\,dx \tag{3.46}$$

pour $k = 0, 1, \ldots$ On a: $\mu_0 = 1$, $\mu_1 = 0$ et $\mu_2 = \sigma_X^2$.
Exemple 3.5.2 (suite). Dans le cas où $X \sim \mathrm{Exp}(\lambda)$, on calcule (voir (3.36))

$$E[X^k] = \int_0^{\infty} x^k \lambda e^{-\lambda x}\,dx = \lambda \frac{k!}{\lambda^{k+1}} = \frac{k!}{\lambda^k}$$

♢

Remarque. On peut montrer, en utilisant la *formule du binôme de Newton*, que

$$\mu_k \equiv E[(X-\mu_X)^k] = \sum_{i=0}^{k}(-1)^i \binom{k}{i} \mu'_{k-i}\mu_X^i \tag{3.47}$$

Cette formule nous permet de vérifier que $\mathrm{VAR}[X] = E[X^2] - (E[X])^2$, ce qui peut être réécrit comme suit:

$$\mu_2 = \mu'_2 - \mu_X^2 = \mu'_2 - (\mu'_1)^2 \tag{3.48}$$

Deux autres quantités qui sont utilisées pour caractériser la distribution d'une variable aléatoire X sont le *coefficient d'asymétrie* et le *coefficient d'aplatissement*. Ces deux coefficients sont définis en fonction des moments de X.

D'abord, la quantité $\mu_3 \equiv E[(X-\mu_X)^3]$ est utilisée pour mesurer le degré d'**asymétrie** des distributions de probabilité. Si la distribution est symétrique par rapport à sa moyenne μ_X, alors $\mu_3 = 0$ (si l'on suppose que μ_3 existe). Si $\mu_3 > 0$ (respectivement < 0), alors la distribution est dite asymétrique **vers la droite** (respectivement **vers la gauche**).

Remarque. En fait, si la distribution est symétrique par rapport à sa moyenne et si tous ses moments centrés existent, alors on peut écrire que $\mu_{2k+1} = 0$ pour $k = 0, 1, \ldots$

Définition 3.5.9. *Le* **coefficient d'asymétrie** *d'une variable aléatoire X est défini par*

$$\beta_1 = \frac{\mu_3^2}{\sigma^6} \tag{3.49}$$

Remarques.
(i) Le coefficient β_1 est une quantité sans unité de mesure.
(ii) Certains auteurs travaillent plutôt avec le coefficient $\gamma_1 := \sqrt{\beta_1}$.

Exemple 3.5.2 (suite). On peut écrire que

$$\mu_3 \equiv E[(X-\mu_X)^3] = E[X^3 - 3\mu_X X^2 + 3\mu_X^2 X - \mu_X^3]$$

Comme nous le verrons au chapitre 4, l'espérance mathématique d'une *combinaison linéaire* de variables aléatoires peut être obtenue en remplaçant chaque variable par sa moyenne (par linéarité de l'espérance). Il s'ensuit que

$$\mu_3 = E[X^3] - 3\mu_X E[X^2] + 3\mu_X^2 E[X] - \mu_X^3$$

En utilisant la formule $E[X^k] = k!/\lambda^k$, on obtient:

$$\mu_3 = \frac{3!}{\lambda^3} - \frac{3}{\lambda}\frac{2!}{\lambda^2} + \frac{3}{\lambda^2}\frac{1!}{\lambda} - \frac{1}{\lambda^3} = \frac{2}{\lambda^3}$$

Ainsi, on trouve que

$$\beta_1 = \frac{(2/\lambda^3)^2}{(1/\lambda^2)^3} = 4$$

Donc, toutes les distributions exponentielles possèdent le même coefficient d'asymétrie β_1, peu importe la valeur du paramètre λ. Cela reflète le fait qu'elles ont toutes la même forme (voir la figure 3.9). ♢

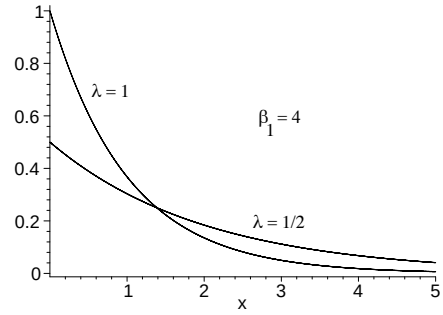

Fig. 3.9. Coefficient d'asymétrie des distributions exponentielles

Exemple 3.5.3. Soit $X \sim \mathrm{U}(a, b)$; puisque la fonction de densité $f_X(x)$ est constante, elle est symétrique par rapport à $\mu_X = (a + b)/2$. Étant donné que tous les moments de la variable aléatoire X existent (car elle est bornée par a et b), il s'ensuit que $\beta_1 = 0$ (voir la figure 3.10). ♢

Définition 3.5.10. *Le* **coefficient d'aplatissement** *d'une variable aléatoire X est la quantité sans unité de mesure*

$$\beta_2 = \frac{\mu_4}{\sigma^4} \tag{3.50}$$

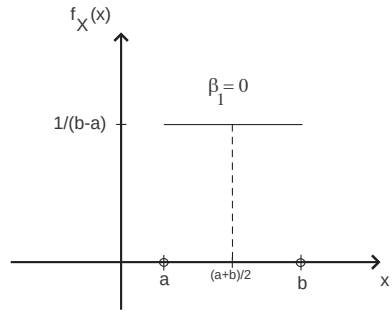

Fig. 3.10. Coefficient d'asymétrie des distributions uniformes

Remarques.
(i) Lorsque la distribution de X est symétrique, β_2 est une mesure de l'*épaisseur* relative des extrémités de la distribution par rapport à sa partie centrale.

(ii) Comme dans le cas du coefficient β_1, certains auteurs utilisent un coefficient différent: $\gamma_2 := \beta_2 - 3$. Puisque $\beta_2 = 3$ si $X \sim \mathrm{N}(\mu, \sigma^2)$, la quantité γ_2 est choisie de façon à obtenir un coefficient d'aplatissement nul pour les distributions normales.

Exemple 3.5.2 (suite). En utilisant encore une fois la formule $E[X^k] = k!/\lambda^k$, on peut écrire que

$$\begin{aligned}\mu_4 \equiv E[(X-\mu_X)^4] &= E[X^4 - 4\mu_X X^3 + 6\mu_X^2 X^2 - 4\mu_X^3 X + \mu_X^4] \\ &= E[X^4] - 4\mu_X E[X^3] + 6\mu_X^2 E[X^2] - 4\mu_X^3 E[X] + \mu_X^4 \\ &= \frac{4!}{\lambda^4} - \frac{4}{\lambda}\frac{3!}{\lambda^3} + \frac{6}{\lambda^2}\frac{2!}{\lambda^2} - \frac{4}{\lambda^3}\frac{1}{\lambda} + \frac{1}{\lambda^4} = \frac{9}{\lambda^4}\end{aligned}$$

Donc, on a:

$$\beta_2 = \frac{9/\lambda^4}{(1/\lambda^2)^2} = 9$$

ce qui est indépendant du paramètre λ. ♢

Exemple 3.5.4. Comme nous l'avons mentionné ci-dessus, on peut montrer que si $X \sim \mathrm{N}(\mu, \sigma^2)$, alors $\beta_2 = 3$. Une variable aléatoire dont la fonction de densité ressemble beaucoup à celle d'une distribution normale est la *distribution de Student* à n degrés de liberté, que nous verrons au chapitre 5. Son coefficient d'aplatissement est donné par

$$\beta_2 = \frac{6}{n-4} + 3 \quad \text{si } n > 4$$

On voit que le coefficient β_2 est supérieur à celui des distributions normales, reflétant le fait que la fonction de densité $f_X(x)$ tend moins rapidement vers 0 lorsque x tend vers $\pm\infty$ que pour les distributions normales. Cependant, on voit aussi que β_2 décroît vers 3 lorsque n tend vers l'infini. ♢

Exemple 3.5.5. Soit $X \sim \mathrm{B}(1,p)$; c'est-à-dire que X présente une distribution de Bernoulli de paramètre p. On a:

$$\mu'_k \equiv E[X^k] = \sum_{x=0}^{1} x^k p_X(x) = 0^k q + 1^k p = p$$

pour $k = 1, 2, \ldots$ En particulier, $E[X] = \mu'_1 = p$, de sorte que

$$\mu_k = E[(X-p)^k] = \sum_{x=0}^{1} (x-p)^k p_X(x) = (-p)^k q + (1-p)^k p$$

Il s'ensuit que $\mathrm{VAR}[X] = \mu_2 = p^2 q + q^2 p = pq(p+q) = pq$. On a aussi:

$$\mu_3 = -p^3 q + q^3 p = pq(-p^2 + q^2)$$

et

$$\mu_4 = p^4 q + q^4 p = pq(p^3 + q^3)$$

Alors on calcule

$$\beta_1 \equiv \frac{\mu_3^2}{\sigma^6} = \frac{p^2 q^2 (-p^2+q^2)^2}{(pq)^3} = \frac{p^3}{q} + \frac{q^3}{p} - 2pq$$

et

$$\beta_2 \equiv \frac{\mu_4}{\sigma^4} = \frac{pq(p^3+q^3)}{(pq)^2} = \frac{p^2}{q} + \frac{q^2}{p}$$

♢

Remarque. Les coefficients β_1 et β_2 sont souvent utilisés pour comparer une distribution quelconque à une distribution normale, pour laquelle $\beta_1 = 0$ (par symétrie) et $\beta_2 = 3$. Par exemple, si X présente une distribution du khi-deux à n degrés de liberté (voir la distribution gamma ci-dessus), alors $\beta_1 = 8/n$ et

$$\beta_2 = 3\left(\frac{4}{n} + 1\right)$$

Notons que β_1 décroît vers 0 et β_2 vers 3 lorsque n tend vers l'infini, comme pour une distribution normale. Cela est une conséquence du *théorème central limite* (chapitre 4).

Nous allons terminer cette section en donnant une proposition qui permet d'obtenir une borne pour une certaine probabilité, lorsque la moyenne et la variance de la variable aléatoire sont finies (et connues).

Proposition 3.5.1. (Inégalité de Bienaymé-Tchebychev) *Pour toute constante $a > 0$, on a:*

$$P[\mu_X - a\sigma_X \leq X \leq \mu_X + a\sigma_X] > 1 - \frac{1}{a^2} \tag{3.51}$$

pour toute variable aléatoire X dont la variance $VAR[X] = \sigma_X^2$ est finie.

Remarques.
(i) Pour que la variance de X soit finie, il faut que sa moyenne $E[X]$ aussi soit finie.

(ii) Généralement, on dit que la moyenne (respectivement la variance) d'une variable aléatoire X n'existe pas si $E[X] = \pm\infty$ (respectivement $\mathrm{VAR}[X] = \infty$). C'est pourquoi dans plusieurs livres la condition de validité de l'inégalité de Bienaymé-Tchebychev est que la variance de X existe. On peut toutefois faire une distinction entre le cas où la moyenne (ou la variance) de X est infinie et celui où elle n'existe pas. Ainsi, la moyenne d'une *distribution de Cauchy* (voir la page 9) n'existe pas, car on trouve que $E[X] = \infty - \infty$, ce qui n'est pas défini. Alors sa variance non plus n'existe pas.

Exemple 3.5.6. Si X est une variable aléatoire pour laquelle $E[X] = 0$ et $\mathrm{VAR}[X] = 1$, alors on peut écrire que

$$P[-3 \leq X \leq 3] > 1 - \frac{1}{9} = 0{,}\bar{8}$$

Dans le cas d'une distribution $\mathrm{N}(0, 1)$, cette probabilité est en fait supérieure à 99,7 % (tel que mentionné ci-dessus). Si X présente une distribution uniforme sur l'intervalle $[-\sqrt{3}, \sqrt{3}]$, de sorte que sa moyenne est nulle et sa variance est égale à 1, alors la probabilité en question est égale à 1 (puisque $[-\sqrt{3}, \sqrt{3}] \subset [-3, 3]$). ♢

3.6 Exercices du chapitre 3

Exercices résolus

Question n° 1

Soit

$$p_X(x) = \begin{cases} a/8 \text{ si } x = -1 \\ a/4 \text{ si } x = 0 \\ a/8 \text{ si } x = 1 \end{cases}$$

où $a > 0$. Trouver la constante a.

Solution. On doit avoir: $1 = \frac{a}{8} + \frac{a}{4} + \frac{a}{8} = \frac{a}{2}$. Alors il faut que $a = 2$.

Question n° 2

Soit

$$f_X(x) = \frac{3}{4}\left(1 - x^2\right) \quad \text{si } -1 < x < 1$$

Calculer $F_X(0)$.

Solution. On calcule

$$F_X(0) = \int_{-1}^{0} \tfrac{3}{4}(1 - x^2)\, dx = \frac{3}{4}\left(x - \frac{x^3}{3}\right)\Big|_{-1}^{0} = \frac{1}{2}$$

Remarque. On pouvait déduire ce résultat de la symétrie de la fonction f_X par rapport à 0.

Question n° 3

Calculer l'écart-type de X si

$$p_X(x) = \frac{1}{3} \quad \text{pour } x = 1, 2 \text{ ou } 3$$

Solution. On calcule d'abord $E[X] = \frac{1}{3}(1 + 2 + 3) = 2$. Ensuite, on a:

$$E[X^2] = \frac{1}{3}(1^2 + 2^2 + 3^2) = \frac{14}{3} \implies \text{VAR}[X] = \frac{14}{3} - 2^2 = \frac{2}{3}$$

Donc, on peut écrire que $\text{STD}[X] = \sqrt{2/3}$.

Question n$^{\text{o}}$ 4

Supposons que

$$f_X(x) = 2x \quad \text{si } 0 < x < 1$$

Calculer $E\left[X^{1/2}\right]$.

Solution. On a:

$$E\left[X^{1/2}\right] = \int_0^1 x^{1/2}\, 2x\, dx = 2\,\frac{x^{5/2}}{5/2}\bigg|_0^1 = \frac{4}{5}$$

Question n$^{\text{o}}$ 5

Calculer le 25$^{\text{e}}$ centile de la variable aléatoire X pour laquelle

$$F_X(x) = \frac{x^2}{4} \quad \text{si } 0 < x < 2$$

Solution. On peut écrire que $x_{0,75}^2/4 = 0{,}25$ et $x_{0,75} > 0$. Il s'ensuit que $x_{0,75} = 1$.

Question n$^{\text{o}}$ 6

Soit

$$f_X(x) = \frac{1}{2} \quad \text{si } 0 < x < 2$$

On définit $Y = X + 1$. Calculer $f_Y(y)$.

Solution. On a:

$$f_Y(y) = f_X(y-1)\left|\frac{d}{dy}(y-1)\right| = \frac{1}{2} \quad \text{si } 1 < y < 3$$

Question n$^{\text{o}}$ 7

Deux objets sont pris au hasard et sans remise dans une boîte qui contient 5 objets de marque A et 10 objets de marque B. Soit X le nombre d'objets de marque A parmi les deux objets choisis. Quelle est la distribution de probabilité de X et ses paramètres?

Solution. Par définition, $X \sim \text{Hyp}(N = 15, n = 2, d = 5)$.

Question n$^{\text{o}}$ 8

Supposons que $X \sim \text{B}(n = 5, p = 0{,}25)$. Quel est le mode, c'est-à-dire la valeur la plus probable, de X?

Solution. D'après le tableau A.1, page 514, $x = 1$ est la valeur la plus probable, avec $p_X(1) \simeq 0{,}6328 - 0{,}2373 = 0{,}3955$.

Question n° 9

Dix pour cent des articles produits par une certaine machine sont défectueux. Si l'on prend au hasard dix articles (indépendants) fabriqués par cette machine, quelle est la probabilité qu'exactement deux de ceux-ci soient défectueux?

Solution. On cherche $P[\mathrm{B}(10; 0{,}1) = 2] \overset{\text{tab. A.1}}{\simeq} 0{,}9298 - 0{,}7361 = 0{,}1937$.

Question n° 10

Calculer $P\left[X \geq 1 \mid X \leq 1\right]$ si $X \sim \mathrm{Poi}(\lambda = 5)$.

Solution. On a:

$$P[X \geq 1 \mid X \leq 1] = \frac{P[X = 1]}{P[X \leq 1]} = \frac{e^{-5} \cdot 5}{e^{-5} + e^{-5} \cdot 5} = \frac{5}{6}$$

Question n° 11

Des pannes se produisent selon un processus de Poisson de taux $\lambda = 2$ par jour. Calculer la probabilité qu'au cours de deux journées consécutives il se produise au total exactement une panne.

Solution. On cherche $P[\mathrm{Poi}(2 \cdot 2) = 1] = e^{-4} \cdot 4 \simeq 0{,}0733$.

Question n° 12

Dans un lac, il y a 200 poissons de type I et 50 poissons de type II. On prend, sans remise, cinq poissons dans le lac. Utiliser la distribution binomiale pour calculer approximativement la probabilité que l'on n'ait aucun poisson de type II.

Solution. On peut écrire que $P[\mathrm{Hyp}(N = 250, n = 5, d = 50) = 0]$

$$\simeq P\left[\mathrm{B}\left(n = 5, p = 50/250\right) = 0\right] = \left(\frac{4}{5}\right)^5 \simeq 0{,}3277$$

Question n° 13

Soit $X \sim \mathrm{B}(n = 50, p = 0{,}01)$. Utiliser une distribution de Poisson pour calculer approximativement $P\left[X \geq 4\right]$.

Solution. On a:

$$P[\mathrm{B}(50;0{,}01) \geq 4] \quad \simeq \quad P[\mathrm{Poi}(1/2) \geq 4] = 1 - P[\mathrm{Poi}(1/2) \leq 3]$$
$$\overset{\text{tab. A.2}}{\simeq} 1 - 0{,}9982 = 0{,}0018$$

Question n$^{\mathrm{o}}$ 14

On lance une pièce de monnaie bien équilibrée jusqu'à ce que l'on obtienne "face". Quelle est la probabilité que l'expérience se termine lors du cinquième lancer?

Solution. On cherche

$$P[\mathrm{Géom}(p = 1/2) = 5] = \left(\frac{1}{2}\right)^{5-1}\left(\frac{1}{2}\right) = \frac{1}{32} = 0{,}03125$$

Question n$^{\mathrm{o}}$ 15

La durée de vie X d'un appareil de radio présente une distribution exponentielle de moyenne égale à 10 ans. Quelle est la probabilité qu'un appareil vieux de 10 ans fonctionne encore au bout de 10 années additionnelles?

Solution. Par la propriété de non-vieillissement de la distribution exponentielle, on peut écrire que

$$P[X > 20 \mid X > 10] = P[X > 10] = P[\mathrm{Exp}(\lambda = 1/10) > 10] = e^{-1} \simeq 0{,}3679$$

Question n$^{\mathrm{o}}$ 16

Supposons que $X \sim \mathrm{Exp}(\lambda)$. Trouver λ tel que $P\,[X > 1] = 2P\,[X > 2]$.

Solution. On a:

$$P[X > 1] = 2\,P[X > 2] \iff e^{-\lambda} = 2\,e^{-2\lambda} \iff e^{\lambda} = 2 \iff \lambda = \ln 2$$

Question n$^{\mathrm{o}}$ 17

Soit $X \sim \mathrm{G}(\alpha = 2, \lambda = 1)$. Quelle autre distribution de probabilité peut-on utiliser pour calculer *exactement* $P\,[X < 4]$? Donner aussi le ou les paramètres de cette distribution.

Solution. On peut écrire que

$$P[\mathrm{G}(\alpha = 2, \lambda = 1) < 4] = P[\mathrm{Poi}(4 \cdot 1) \geq 2]$$

Donc, on peut utiliser la distribution de Poisson de paramètre 4.

Question n^{o} 18

Des clients se présentent à un vendeur selon un processus de Poisson, au rythme moyen de deux clients par heure. Quelle est la distribution du temps X requis pour que 10 clients se présentent? Donner aussi le ou les paramètres de cette distribution.

Solution. On peut écrire que $X \sim \mathrm{G}(\alpha = 10, \lambda = 2)$.

Question n^{o} 19

Calculer $P\left[|X| < 1/2\right]$ si $X \sim \mathrm{N}(0,1)$.

Solution. On a:

$$P[|\mathrm{N}(0,1)| < 1/2] \overset{\text{sym.}}{=} 2\Phi(1/2) - 1 \overset{\text{tab. A.3}}{\simeq} 2(0{,}6915) - 1 = 0{,}3830$$

Question n^{o} 20

Calculer le 10^{e} centile de $X \sim \mathrm{N}(1,2)$.

Solution. On a:

$$x_{0,90} = \mu_X + z_{0,90} \cdot \sigma_X \overset{\text{tab. A.4}}{\simeq} 1 + \sqrt{2}(-1{,}282) \simeq -0{,}813$$

Question n^{o} 21

Des appareils sont constitués de cinq composants indépendants. Un appareil donné fonctionne si au moins quatre de ses cinq composants fonctionnent. Chaque composant fonctionne avec une probabilité de 0,95. On reçoit un très grand lot de ces appareils et on examine un appareil à la fois, pris au hasard et avec remise, jusqu'à ce que l'on ait obtenu un appareil défectueux.

(a) Quelle est la probabilité qu'un appareil pris au hasard fonctionne?

(b) Quelle est l'espérance du nombre d'appareils qu'il faudra examiner?

Solution. (a) Soit X le nombre de composants en panne; alors on a: $X \sim \mathrm{B}(n = 5, p = 0{,}05)$. On cherche $P[X \leq 1] \overset{\text{tab. A.2}}{\simeq} 0{,}977$.

(b) Soit Y le nombre d'appareils requis pour obtenir un premier appareil défectueux. On a: $Y \sim \mathrm{Géom}(p = P[X > 1] \simeq 1 - 0{,}977)$. Alors $E[Y] = 1/p \simeq 43{,}5$ appareils.

Question n° 22

Les autobus de la ville passent à un certain coin de rue, entre 7 h et 19 h 30, selon un processus de Poisson à la cadence moyenne de quatre par heure.

(a) Quelle est la probabilité qu'au moins 30 minutes s'écoulent entre le passage du premier et du troisième autobus?

(b) Quelle est la variance du temps d'attente entre le premier et le troisième autobus?

(c) Étant donné qu'une personne attend l'autobus depuis 5 minutes, quelle est la probabilité qu'elle attende encore pendant 15 minutes?

Solution. (a) Soit $N(t)$ le nombre d'autobus qui passent dans un intervalle de t heures. Alors $N(t) \sim \text{Poi}(4\,t)$. On cherche

$$P[N(1/2) \leq 1] = P[\text{Poi}(2) \leq 1] = e^{-2} + 2e^{-2} \simeq 0{,}4060$$

(b) Soit T le temps d'attente entre le premier et le troisième autobus. Alors $T \sim \text{G}(\alpha = 2, \lambda = 4)$, de sorte que $\text{VAR}[T] = \alpha/\lambda^2 = 1/8$ (heure)^2.

(c) Soit W le temps d'attente en minutes. Alors $W \sim \text{Exp}(1/15)$. On cherche

$$P[W > 20 \mid W > 5] = P[W > 15] = e^{-15/15} \simeq 0{,}3679$$

Question n° 23

On suppose que la longueur X (en mètres) des places de stationnement présente une distribution $\text{N}(\mu; 0{,}01\ \mu^2)$.

(a) Vous avez une voiture de luxe, dont la longueur est 15 % plus grande que la longueur moyenne d'une place de stationnement. Quelle proportion des places de stationnement libres vous est accessible?

(b) Si $\mu = 4$, quelle aurait dû être la longueur de votre automobile pour avoir accès à 90 % des places de stationnement libres?

Solution. (a) On calcule

$$\begin{aligned} P[\text{N}(\mu; (0{,}1\mu)^2) \geq 1{,}15\mu] &= 1 - \Phi\left(\frac{1{,}15\mu - \mu}{0{,}1\mu}\right) \\ &= 1 - \Phi(1{,}5) \overset{\text{tab. A.3}}{\simeq} 1 - 0{,}9332 = 0{,}0668 \end{aligned}$$

(b) On cherche $x_{0,10} = \mu + z_{0,10} \cdot \sigma \overset{\text{tab. A.4}}{\simeq} 4 - (0{,}1)(4)(1{,}282) \simeq 3{,}49$.

Question n° 24

Une étudiante peut résoudre, en moyenne, la moitié de ses problèmes de probabilités. Dans un examen, il y a 10 questions indépendantes. Quelle est la probabilité qu'elle en réussisse plus de la moitié?

Solution. Soit X le nombre de questions réussies. Alors X présente une distribution $\mathrm{B}(n = 10, p = 0{,}5)$. On cherche

$$P[X > 5] = 1 - P[X \leq 5] \overset{\text{tab. A.1}}{\simeq} 1 - 0{,}6230 = 0{,}3770$$

Question n° 25

Soit X une variable aléatoire qui présente une distribution binomiale de paramètres $n = 100$ et $p = 0{,}1$. Utiliser une distribution de Poisson pour calculer approximativement $P[X = 15]$.

Solution. On a:

$$P[\mathrm{B}(100; 0{,}1) = 15] \simeq P[\mathrm{Poi}(10) = 15] = e^{-10}\frac{10^{15}}{15!} \simeq 0{,}0347$$

Question n° 26

Soit

$$f_X(x) = \sqrt{\frac{2}{\pi} - x^2} \quad \text{pour } -\sqrt{\tfrac{2}{\pi}} \leq x \leq \sqrt{\tfrac{2}{\pi}}$$

Calculer $P[X < 0]$.

Solution. On peut effectuer l'intégrale requise pour obtenir la probabilité demandée. Cependant, on remarque que la fonction $f_X(x)$ est symétrique par rapport à 0; c'est-à-dire que $f_X(-x) = f_X(x)$. Alors, étant donné que X est une variable aléatoire continue qui est définie dans un intervalle borné, on peut écrire que $P[X < 0] = 1/2$.

Question n° 27

Les résultats d'un test d'intelligence pour les élèves d'une certaine école élémentaire ont montré que le quotient intellectuel (QI) de ces élèves présente (approximativement) une distribution normale de paramètres $\mu = 100$ et $\sigma^2 = 225$. Quel pourcentage total des élèves a un QI inférieur à 91 ou supérieur à 130?

Solution. Soit X le QI des élèves. Alors $X \sim \mathrm{N}(100, (15)^2)$. On cherche

$$P[\{X < 91\} \cup \{X > 130\}] = 1 - P\left[\frac{91-100}{15} \leq \mathrm{N}(0,1) \leq \frac{130-100}{15}\right]$$
$$= 1 - [\Phi(2) - \Phi(-0{,}6)] \overset{\text{tab. A.3}}{\simeq} 1 - [0{,}9772 - (1 - 0{,}7257)] = 0{,}2971$$

Question n° 28

Un certain montage nécessite 59 transistors en bon état. On dispose de 60 transistors pris au hasard dans une production en série, dont on sait par expérience qu'elle donne 5 % de déchets (transistors en mauvais état).

(a) Calculer la probabilité qu'on puisse faire ce montage.

(b) Obtenir une valeur approximative de cette probabilité à l'aide d'une distribution de Poisson.

(c) Si l'on dispose en fait d'un très grand nombre de transistors, dont 5 % sont en mauvais état, et si l'on en prend au hasard jusqu'à ce qu'on en ait 59 en bon état, quelle est la probabilité qu'on doive en prendre exactement 60?

Solution. (a) Soit X le nombre de transistors en *mauvais* état. Alors X présente une distribution $\mathrm{B}(n = 60, p = 0{,}05)$. On cherche

$$P[X \leq 1] = (0{,}95)^{60} + \binom{60}{1}(0{,}05)(0{,}95)^{59} = (0{,}95)^{59}(0{,}95 + 3) \simeq 0{,}1916$$

(b) On peut écrire que

$$P[X \leq 1] \simeq P[\mathrm{Poi}(60 \times 0{,}05 = 3) \leq 1] = e^{-3}(1 + 3) \simeq 0{,}1991$$

(c) Soit W le nombre de transistors que l'on doit prendre pour obtenir les 59 transistors en bon état; alors $W \sim \mathrm{BN}(r = 59, p = 0{,}95)$. On cherche

$$P[W = 60] = \binom{59}{1}(0{,}95)^{59}(0{,}05) = 2{,}95(0{,}95)^{59} \simeq 0{,}1431$$

Question n° 29

Soit X le délai de livraison (en jours) d'un produit donné. On sait que X est une variable aléatoire continue dont la moyenne est égale à 7 et l'écart-type est égal à 1. Déterminer un intervalle de temps qui assure, quelle que soit la distribution de X, que les délais de livraison seront dans cet intervalle avec une probabilité d'au moins 90 %.

Solution. Par l'inégalité de Bienaymé-Tchebychev, on peut écrire que

$$P[7 - k \cdot 1 \leq X \leq 7 + k \cdot 1] \geq 1 - \frac{1}{k^2}$$

On veut que $1 - \frac{1}{k^2} = 0{,}90$. On en déduit que $k = \sqrt{10}$. De là, l'intervalle recherché est $[7 - \sqrt{10}, 7 + \sqrt{10}] \simeq [3{,}84; 10{,}16]$.

Question n^o 30

On définit l'entropie H d'une variable aléatoire continue X par $H = E\left[-\ln f_X(X)\right]$, où f_X est la fonction de densité de X et ln désigne le logarithme naturel. Calculer l'entropie d'une variable aléatoire normale de moyenne nulle et de variance $\sigma^2 = 2$.

Solution. On a:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sqrt{2}} \exp\left\{-\frac{x^2}{2(2)}\right\} \implies \ln[f_X(X)] = \ln\left(\frac{1}{2\sqrt{\pi}}\right) - \frac{X^2}{4}$$

Puisque $E[X^2] = \text{VAR}[X] + (E[X])^2 = 2 + 0^2 = 2$, on peut alors écrire que

$$\begin{aligned} H = E\left[-\ln\left(\frac{1}{2\sqrt{\pi}}\right) + \frac{X^2}{4}\right] &= \ln(2\sqrt{\pi}) + \frac{1}{4}E[X^2] \\ &= \ln(2\sqrt{\pi}) + \frac{2}{4} \simeq 1{,}766 \end{aligned}$$

Question n^o 31

Le nombre N d'appareils qu'un technicien doit tenter de réparer au cours d'une journée quelconque est une variable aléatoire qui présente une distribution géométrique de paramètre $p = 1/8$. On estime que la probabilité qu'il réussisse à réparer un appareil donné est égale à 0,95, indépendamment d'un appareil à l'autre.

(a) Quelle est la probabilité que le technicien réussisse à réparer exactement cinq appareils, avant d'être incapable de réparer un deuxième appareil au cours d'une journée donnée, si l'on suppose qu'il recevra au moins sept appareils à réparer durant cette même journée?

(b) Si, au cours d'une journée donnée, le technicien a eu exactement 10 appareils à réparer, quelle est la probabilité qu'il ait réussi à en réparer exactement 8?

(c) Utiliser une distribution de Poisson pour calculer approximativement la probabilité en (b).

(d) Supposons que 8 des 10 appareils en (b) ont effectivement été réparés. Si l'on prend 3 appareils au hasard et sans remise parmi les 10 qu'il avait à réparer, quelle est la probabilité que les 2 appareils non réparés soient parmi ceux-ci?

Solution. (a) Soit X le nombre total d'appareils que le technicien aura tenté de réparer au moment de son deuxième échec. Alors $X \sim \text{BN}(r = 2, p = 0{,}05)$. On cherche

$$P[X = 7] = \binom{6}{1}(0{,}05)^2(0{,}95)^5 \simeq 0{,}0116$$

Remarque. Si l'on ne nous avait pas mentionné que le technicien aura au moins sept appareils à réparer cette journée-là, il aurait alors fallu multiplier la probabilité ci-dessus par celle de recevoir au moins sept appareils au cours d'une journée, soit $P[N \geq 7] = P[N > 6] = (7/8)^6 \simeq 0{,}4488$.

(b) On cherche

$$P[\text{B}(n = 10, p = 0{,}95) = 8] = \underbrace{\binom{10}{8}}_{45}(0{,}95)^8(0{,}05)^2 \simeq 0{,}0746$$

Remarque. On peut aussi écrire que

$$\begin{aligned} P[\text{B}(n = 10, p = 0{,}95) = 8] &= P[\text{B}(n = 10, p = 0{,}05) = 2] \\ &\overset{\text{tab. A.1}}{\simeq} 0{,}9885 - 0{,}9139 = 0{,}0746 \end{aligned}$$

(c) On a:

$$\begin{aligned} &P[\text{B}(n = 10, p = 0{,}95) = 8] = P[\text{B}(n = 10, p = 0{,}05) = 2] \\ &\simeq P[\text{Poi}(\lambda = 10 \times 0{,}05) = 2] = e^{-0{,}5}\frac{(0{,}5)^2}{2} \simeq 0{,}0758 \end{aligned}$$

Remarque. On pouvait aussi utiliser le tableau A.2, à la page 515, pour obtenir la probabilité $P[\text{Poi}(\lambda = 0{,}5) = 2]$.

(d) Soit M le nombre d'appareils non réparés parmi les trois pris au hasard. Alors $M \sim \text{Hyp}(N = 10, n = 3, d = 2)$. On cherche

$$P[M = 2] = \frac{\binom{2}{2}\binom{8}{1}}{\binom{10}{3}} = \frac{1 \times 8}{\frac{10!}{3!7!}} = \frac{1}{15} \simeq 0{,}0667$$

Question n^{o} 32

Le nombre X de raisins dans un biscuit présente une distribution de Poisson de paramètre λ. Quelle valeur de λ doit-on choisir si l'on veut qu'au plus 2 biscuits, dans un sac de 20, ne contiennent aucun raisin, et ce, avec une probabilité de 92,5 %?

Solution. Soit Y le nombre de biscuits, parmi les 20, qui ne contiennent aucun raisin. Si l'on suppose que le nombre de raisins que contient un biscuit est indépendant du nombre de raisins dans les autres biscuits, alors on peut écrire que $Y \sim \mathrm{B}(n = 20, p = P[X = 0])$. On a: $P[X = 0] = e^{-\lambda}$. De plus, on trouve dans le tableau A.2, à la page 515, que

$$P[Y \leq 2] \simeq 0{,}9245 \quad \text{si } p = 0{,}05$$

Il s'ensuit que l'on doit choisir $\lambda \simeq -\ln 0{,}05 \simeq 3$.

Remarque. On peut vérifier qu'avec $\lambda = 3$, on a:

$$\begin{aligned} P[Y \leq 2] \simeq\ & (0{,}9502)^{20} + 20(0{,}9502)^{19}(0{,}0498) \\ & + \underbrace{\binom{20}{2}}_{190}(0{,}9502)^{18}(0{,}0498)^2 \simeq 0{,}925 \end{aligned}$$

Cependant, il n'était pas nécessaire que λ soit un entier, car il s'agit du nombre *moyen* de raisins dans les biscuits.

Question n^{o} 33

Le réservoir d'une station-service est rempli d'essence une fois par semaine. On suppose que la demande hebdomadaire X (en milliers de litres) est une variable aléatoire qui présente une distribution exponentielle de paramètre $\lambda = 1/10$. Quelle doit être la capacité du réservoir pour que la probabilité d'épuiser l'approvisionnement d'une semaine soit de 0,01?

Solution. Soit c la capacité du réservoir. On veut que $P[X \geq c] = 0{,}01$. On a:

$$P[X \geq c] = \int_c^\infty \frac{1}{10} e^{-x/10}\, dx = e^{-c/10}$$

Donc, il faut que

$$c = -10\,(\ln 0{,}01) \simeq 46 \text{ milliers de litres}$$

Question n^{o} 34

On s'intéresse à la durée de vie X (en années) d'un appareil. Par expérience, on estime à 0,1 la probabilité qu'un appareil de ce type dure plus de neuf ans.

(a) On propose le modèle suivant pour la fonction de densité de X:

$$f_X(x) = \frac{a}{(x+1)^b} \quad \text{pour } x \geq 0$$

où $a > 0$ et $b > 1$. Trouver les constantes a et b.

(b) Si l'on propose une distribution normale de moyenne $\mu = 7$ pour X, quelle doit être la valeur du paramètre σ?

(c) On considère 10 appareils de ce type, que l'on suppose indépendants. Calculer la probabilité que huit ou neuf de ces appareils durent moins de neuf ans.

Solution. (a) On a:

$$1 = \int_0^\infty \frac{a}{(x+1)^b}\,dx \implies \frac{a}{b-1} = 1$$

et alors

$$0{,}1 = \int_9^\infty \frac{b-1}{(x+1)^b}\,dx \implies \frac{1}{10^{b-1}} = 0,1$$

On trouve que $b = 2$, ce qui implique que $a = 1$.

(b) On veut que

$$P[\mathrm{N}(7,\sigma^2) > 9] = 0{,}1 \iff Q\left(\frac{2}{\sigma}\right) = 0{,}1 \iff \frac{2}{\sigma} \overset{\text{tab. A.4}}{\simeq} 1{,}282$$

Donc, il faut que σ soit environ égal à 1,56.

(c) Soit Y le nombre d'appareils, parmi les 10, qui dureront moins de 9 ans. Alors $Y \sim \mathrm{B}(n = 10, p = 0{,}9)$. On cherche

$$\begin{aligned} P[Y \in \{8,9\}] &= P[\mathrm{B}(n = 10, p = 0{,}1) \in \{1,2\}] \\ &\overset{\text{tab. A.1}}{\simeq} 0{,}9298 - 0{,}3487 = 0{,}5811 \end{aligned}$$

Exercices

Question n° 1

Une variable aléatoire continue a pour fonction de densité

$$f_X(x) = \begin{cases} cxe^{-x/2} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

(a) Calculer la constante c.

(b) Déterminer la fonction de répartition $F_X(x)$ (intégrer par parties).

(c) Calculer la moyenne de X.

(d) Calculer l'écart-type de X.

(e) Montrer que la médiane de X est située entre 3 et 4.

Indication. On a:

$$\int_0^\infty x^k e^{-\beta x} dx = \frac{k!}{\beta^{k+1}} \quad \text{pour } k = 0, 1, 2, \ldots$$

Question n° 2

Un commerçant reçoit un lot de 100 appareils électriques. Pour gagner du temps, il décide d'utiliser le plan d'échantillonnage suivant: il prend deux appareils, au hasard et sans remise, et il décide d'accepter le lot si les deux appareils choisis ne sont pas défectueux. Soit X la variable aléatoire qui désigne le nombre d'appareils défectueux dans l'échantillon.

(a) Donner la distribution de probabilité de X ainsi que les paramètres de cette distribution.

(b) Si le lot contient deux appareils défectueux, calculer la probabilité que le lot soit accepté.

(c) Approcher la probabilité calculée en (b) au moyen d'une distribution binomiale.

(d) Approcher la probabilité calculée en (c) en utilisant une distribution de Poisson.

Remarque. Donner vos réponses avec quatre décimales.

Question n° 3

Dans le jeu de fléchettes, on vise une cible circulaire de rayon égal à 25 centimètres. Notons par X la distance (en centimètres) entre le point d'impact de la fléchette et le centre de la cible. On suppose que

$$P[X \leq x] = \begin{cases} c\pi x^2 \text{ si } 0 \leq x \leq 25 \\ 1 \quad \text{si } x > 25 \end{cases}$$

où c est une constante.

(a) Calculer

(i) la constante c;

(ii) la fonction de densité, $f_X(x)$, de X;

(iii) la moyenne de X;

(iv) la probabilité $P[X \leq 10 \mid X \geq 5]$.

(b) Cela coûte 1 \$ pour lancer une fléchette et on gagne

$$\begin{cases} 10\,\$ \text{ si } X \leq r \\ 1\,\$ \text{ si } r < X \leq 2r \\ 0\,\$ \text{ si } 2r < X \leq 25 \end{cases}$$

Pour quelle valeur de r le gain net moyen du joueur est-il de 0,25 \$?

Question n^{o} 4

Des appels à un central téléphonique arrivent selon un processus de Poisson de taux λ par minute. On sait, par expérience, que la probabilité de recevoir exactement un appel durant une minute est égale au triple de la probabilité de ne recevoir aucun appel durant la même période. Pour chacune des questions qui suivent, donner la distribution de probabilité de la variable aléatoire et calculer la probabilité demandée, s'il y a lieu.

(a) Soit X le nombre d'appels reçus durant une minute. Quelle est la probabilité $P[2 \leq X \leq 4]$?

(b) Soit Y le nombre d'appels reçus durant une période de trois minutes. Calculer $P[Y \geq 4]$.

(c) Soit W_1 le temps d'attente (en minutes) jusqu'au premier appel à partir de l'instant $t = 0$. Calculer $P[W_1 \leq 1]$.

(d) Soit W_2 le temps d'attente (en minutes) entre le premier et le deuxième appel. Calculer $P[W_2 > 1]$.

(e) Soit W le temps d'attente jusqu'au deuxième appel à partir de l'instant $t = 0$. Donner la distribution de probabilité de W ainsi que ses paramètres.

(f) On considère 100 périodes consécutives d'une minute et on désigne par U le nombre de ces périodes où aucun appel n'a été reçu. Calculer $P[U \leq 1]$.

Question n° 5

On dispose de 10 machines (indépendantes) produisant chacune 2 % d'articles défectueux.

(a) Combien d'articles seront fabriqués par la première machine, en moyenne, *avant* qu'elle ne produise un premier article défectueux?

(b) On prend au hasard un article produit par chacune des 10 machines. Quelle est la probabilité qu'au plus deux articles parmi ceux obtenus soient défectueux?

(c) Refaire la partie (b) en utilisant une approximation par une distribution de Poisson.

(d) Combien d'articles fabriqués par la première machine doit-on prendre, au minimum, pour que la probabilité d'obtenir au moins un article défectueux soit supérieure à 1/2? On supposera que les articles sont indépendants les uns des autres.

Question n° 6

On se propose d'étudier la proportion θ d'articles non conformes aux spécifications techniques dans un lot d'articles manufacturés. On décide de prendre, au hasard et avec remise, un échantillon de 20 articles dans le lot.

(a) On désigne par X le nombre d'articles non conformes dans l'échantillon.

(i) Donner la fonction de probabilité $p_X(x)$.

(ii) Donner la fonction de probabilité $p_X(x)$ si les tirages sont effectués sans remise et si la taille du lot est de 1000 articles.

(b) Si les tirages sont effectués avec remise et si $\theta = 0{,}25$, calculer

(i) $P[X = 10]$;

(ii) $P[X \geq 10]$ en utilisant une approximation basée sur une distribution de Poisson.

Question n° 7

Soit X une variable aléatoire qui possède la fonction de densité

$$f_X(x) = \begin{cases} c\left(1 - x^2\right) & \text{si } -1 \leq x \leq 1 \\ 0 & \text{si } |x| > 1 \end{cases}$$

où c est une constante positive. Calculer (a) la constante c; (b) la moyenne de X; c) la variance de X; (d) la fonction de répartition $F_X(x)$.

Question n° 8

Le nombre moyen d'articles non conformes produits par un procédé de fabrication est de 6 par période de 25 minutes, selon un processus de Poisson.

On considère une heure de production subdivisée en 12 périodes de 5 minutes. Posons:

X = nombre d'articles non conformes produits durant une période de cinq minutes;

Y = nombre de périodes de cinq minutes requises pour obtenir une première période pendant laquelle aucun article non conforme n'est produit;

Z = nombre de périodes parmi les 12 pendant lesquelles aucun article non conforme n'est produit.

(a) Donner la distribution de X, de Y et de Z ainsi que leur(s) paramètre(s).

(b) À quelle période, en moyenne, aucun article non conforme ne sera produit pour la première fois?

(c) Quelle est la probabilité que, dans exactement 2 des 12 périodes, on n'observe aucun article non conforme?

(d) Quelle est la probabilité que l'on ait produit exactement deux articles non conformes durant une période de cinq minutes, étant donné qu'au plus quatre articles non conformes ont été produits durant cette période?

Question n$^{\text{o}}$ 9

Calculer la variance de $\sqrt{X}$ si

$$p_X(x) = \begin{cases} 1/4 \text{ si } x = 0 \\ 1/2 \text{ si } x = 1 \\ 1/4 \text{ si } x = 2 \end{cases}$$

Question n$^{\text{o}}$ 10

Calculer le 30$^{\text{e}}$ centile de la variable aléatoire continue X dont la fonction de densité est

$$f_X(x) = \begin{cases} x \text{ si } 0 \leq x \leq \sqrt{2} \\ 0 \text{ ailleurs} \end{cases}$$

Question n$^{\text{o}}$ 11

Un livre de 300 pages contient 200 coquilles. Calculer, en utilisant une distribution de Poisson, la probabilité qu'une page particulière comporte au moins deux coquilles.

Question n$^{\text{o}}$ 12

Par expérience, on estime que 85 % des articles produits par une certaine machine sont conformes aux spécifications techniques. Si la machine produit 20

articles par heure, quelle est la probabilité que 8 ou 9 articles fabriqués pendant une période de 30 minutes soient conformes à ces spécifications?

Question n^{o} 13

Calculer $P[X \geq 8]$ si

$$f_X(x) = \begin{cases} \frac{1}{96}x^3 e^{-x/2} & \text{si } x \geq 0 \\ 0 & \text{ailleurs} \end{cases}$$

Question n^{o} 14

La durée de vie d'un certain composant électronique présente une distribution exponentielle de moyenne égale à cinq ans. Sachant que le composant a déjà duré un an, quelle est la probabilité qu'il tombe en panne pendant sa quatrième année de fonctionnement?

Question n^{o} 15

Un système de sécurité est constitué de 10 composants fonctionnant indépendamment les uns des autres. Pour que le système soit opérationnel, il faut qu'au moins cinq composants fonctionnent. Pour vérifier si le système est opérationnel, on examine périodiquement quatre de ses composants pris au hasard (et sans remise). Le système est jugé opérationnel si au moins trois des quatre composants examinés fonctionnent. Si, en fait, seulement 4 des 10 composants fonctionnent, quelle est la probabilité que le système soit jugé opérationnel?

Question n^{o} 16

Calculer le 25^{e} centile d'une variable aléatoire continue X dont la fonction de densité est

$$f_X(x) = \begin{cases} xe^{-x^2/2} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Question n^{o} 17

Calculer la probabilité d'obtenir exactement trois "piles" en lançant 15 fois (de façon indépendante) une pièce de monnaie pour laquelle la probabilité d'obtenir "pile" est de 0,4.

Question n^{o} 18

On tire sans remise un échantillon de 4 articles dans un lot de 10 articles, dont 1 est défectueux. Calculer la probabilité que le nombre d'articles défectueux dans l'échantillon soit égal à 1.

Question n° 19

Des clients se présentent à un comptoir, selon un processus de Poisson, au rythme moyen de cinq par minute. Calculer la probabilité que le nombre de clients soit supérieur ou égal à 10 dans une période de 3 minutes.

Question n° 20

Les arrivées des clients à un comptoir constituent un processus de Poisson de taux $\lambda = 1$ par période de deux minutes. Calculer la probabilité que le temps d'attente (à partir de n'importe quel instant) jusqu'au prochain client soit inférieur à 10 minutes.

Question n° 21

Soit X une variable aléatoire qui présente une distribution $\mathrm{N}(10, 2)$. Trouver son 90^{e} centile.

Question n° 22

Soit

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{x}{2} & \text{si } 0 \leq x \leq 1 \\ \frac{x}{6} + \frac{1}{3} & \text{si } 1 < x < 4 \\ 1 & \text{si } x \geq 4 \end{cases}$$

la fonction de répartition d'une variable aléatoire continue X.

(a) Calculer la fonction de densité de X.

(b) Quel est le 75^{e} centile de X?

(c) Calculer l'espérance mathématique de X.

(d) Calculer $E[1/X]$.

(e) On définit

$$Y = \begin{cases} -1 \text{ si } X \leq 1 \\ 1 \text{ si } X > 1 \end{cases}$$

(i) Trouver $F_Y(0)$.

(ii) Calculer la variance de Y.

Question n° 23

Une boîte contient 100 transistors de marque A et 50 de marque B.

(a) On prend des transistors un par un, au hasard et avec remise, jusqu'à ce qu'on en obtienne un premier de marque B. Quelle est la probabilité que l'on doive prendre 9 ou 10 transistors?

(b) Quel est le nombre minimal de transistors que l'on doit prendre, au hasard et avec remise, afin que la probabilité de n'obtenir que des transistors de marque A soit inférieure à $1/3$?

Question n^{o} 24

Des objets sont fabriqués en série. Pour effectuer un contrôle de la qualité, on prend, chaque heure, 10 objets, au hasard et sans remise, dans une boîte qui en contient 25. Le procédé de fabrication est jugé sous contrôle si au plus un des objets examinés est défectueux.

(a) Si toutes les boîtes examinées contiennent exactement deux objets défectueux, quelle est la probabilité que le procédé de fabrication soit jugé sous contrôle au moins sept fois au cours d'une journée de travail de huit heures?

(b) Utiliser une distribution de Poisson pour évaluer approximativement la probabilité calculée en (a).

(c) Sachant que, lors du dernier contrôle effectué en (a), le procédé de fabrication a été jugé sous contrôle, quelle est la probabilité que l'échantillon de 10 objets ne contenait aucun article défectueux?

Question n^{o} 25

Soit X une variable aléatoire dont la fonction de probabilité est donnée par

x	-1	0	3
$p_X(x)$	0,5	0,2	0,3

(a) Calculer l'écart-type de X.

(b) Calculer l'espérance mathématique de X^3.

(c) Trouver la fonction de répartition de X.

(d) On définit $Y = X^2 + X + 1$; trouver $p_Y(y)$.

Question n^{o} 26

Dans l'usine A, il s'est produit 25 accidents de travail en 2009. Chaque année, l'usine ferme ses portes pendant deux semaines en juillet pour permettre à ses employés de prendre des vacances. Répondre aux questions suivantes en supposant que les accidents de travail se produisent selon un processus de Poisson.

(a) Quelle est la probabilité qu'exactement un des 25 accidents se soit produit au cours des deux premières semaines de 2009?

(b) Si le taux moyen d'accidents reste le même en 2011, quelle est la probabilité qu'il se produise exactement un accident de travail durant les deux premières semaines de cette année-là?

Question n° 27

Dans une certaine loterie, quatre boules sont tirées au hasard et sans remise parmi 20 boules numérotées de 1 à 20. On gagne un prix si la combinaison qu'on a choisie comporte au moins deux bons numéros. Un joueur décide d'acheter un billet par semaine jusqu'à ce qu'il remporte un prix. Quelle est la probabilité qu'il doive acheter moins de dix billets?

Question n° 28

La fonction de densité de la variable aléatoire X est donnée par

$$f_X(x) = \begin{cases} 6x(1-x) \text{ si } 0 \leq x \leq 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

(a) Calculer l'espérance mathématique de $1/X$.

(b) Obtenir la fonction de répartition de X.

(c) On définit

$$Y = \begin{cases} 2 \text{ si } X \geq 1/4 \\ 0 \text{ si } X < 1/4 \end{cases}$$

Calculer $E\left[Y^k\right]$, où k est un entier naturel.

(d) Soit $Z = X^2$; obtenir la fonction de densité de Z.

Question n° 29

La concentration de réactif dans une réaction chimique est une variable aléatoire X dont la fonction de densité est

$$f_X(x) = \begin{cases} 2(1-x) \text{ si } 0 \leq x \leq 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

La quantité de produit final Y (en grammes) est donnée par $Y = 3X$.

(a) Quelle est la probabilité que la concentration de réactif soit égale à 1/2? Justifier.

(b) Calculer la variance de Y.

(c) Obtenir la fonction de densité de Y.

(d) Quelle est la quantité minimale de produit final qui, dans 95 % des cas, n'est pas dépassée?

Question n° 30

Une compagnie d'assurances emploie 20 vendeurs. Chaque vendeur travaille au bureau ou à l'extérieur, et on estime qu'il est au bureau à 14 h 30, lors d'un

jour ouvrable, avec une probabilité de 0,2, et ce, indépendamment des autres journées et des autres vendeurs.

(a) La compagnie veut installer un nombre minimal de bureaux, de sorte qu'un vendeur trouve un bureau de disponible dans au moins 90 % des cas. Trouver ce nombre minimal de bureaux.

(b) Calculer le nombre minimal de bureaux en (a) en utilisant une approximation de Poisson.

(c) Un client a téléphoné au bureau à 14 h 30 lors des deux derniers jours ouvrables, afin de parler à un vendeur particulier. Étant donné qu'il n'a pas encore réussi à joindre le vendeur en question, quelle est la probabilité qu'il doive téléphoner au moins deux autres fois, s'il téléphone toujours à 14 h 30?

Question n° 31

Calculer $\mathrm{VAR}\left[e^X\right]$ si X est une variable aléatoire dont la fonction de probabilité est donnée par

$$p_X(x) = \begin{cases} 1/4 \text{ si } x = 0 \\ 1/4 \text{ si } x = 1 \\ 1/2 \text{ si } x = 4 \\ 0 \quad \text{autrement} \end{cases}$$

Question n° 32

La fonction de densité de la variable aléatoire X est

$$f_X(x) = \begin{cases} -x \text{ si } -1 \leq x \leq 0 \\ x \text{ si } 0 < x \leq 1 \\ 0 \text{ ailleurs} \end{cases}$$

Calculer $F_X(1/2)$.

Question n° 33

Soit

$$f_X(x) = \begin{cases} 1/e \text{ si } 0 < x < e \\ 0 \quad \text{ailleurs} \end{cases}$$

Calculer $f_Y(y)$, où $Y := -2 \ln X$.

Question n° 34

Un lot contient 20 articles, dont 2 sont défectueux. On prend trois articles au hasard et avec remise. Sachant que l'on a obtenu au moins un article défectueux, quelle est la probabilité que l'on ait obtenu trois articles défectueux?

Question n° 35

Des appels arrivent à un central téléphonique, selon un processus de Poisson, au rythme moyen de deux par minute. Quelle est la probabilité que, pendant au moins une des cinq premières minutes d'une heure quelconque, il n'y ait aucun appel?

Question n° 36

Une boîte contient 20 roches de type granite et 5 roches de type basalte. Dix roches sont prises au hasard et sans remise. Utiliser une distribution binomiale pour calculer approximativement la probabilité que l'on obtienne les cinq roches de type basalte dans l'échantillon.

Question n° 37

On lance une pièce de monnaie bien équilibrée jusqu'à ce qu'on ait obtenu dix fois "face". Quelle est la variance du nombre de lancers requis pour terminer l'expérience aléatoire?

Question n° 38

Soit X une variable aléatoire qui présente une distribution exponentielle de paramètre λ. On définit $Y = \text{ent}(X)+1$, où $\text{ent}(X)$ désigne la *partie entière* de X. Calculer $F_Y(y)$.

Question n° 39

Soit

$$f_X(x) = \begin{cases} 4x^2 e^{-2x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

Calculer la variance de X.

Question n° 40

Supposons que $X \sim \text{N}\left(1, \sigma^2\right)$. Trouver σ si $P[-1 < X < 3] = 0{,}5$.

Question n° 41

La fonction de répartition de la variable aléatoire discrète X est donnée par

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1/2 & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

Calculer (a) $p_X(x)$; (b) $E[\cos(\pi X)]$.

Question n° 42

Au moins la moitié des moteurs d'un avion doivent fonctionner pour que celui-ci puisse voler. Si chaque moteur fonctionne, indépendamment des autres,

avec une probabilité p de 0,6, un avion possédant quatre moteurs est-il plus fiable qu'un avion ayant deux moteurs? Justifier votre réponse.

Question n^{o} 43

On définit $Y = |X|$, où X est une variable continue dont la fonction de densité est

$$f_X(x) = \begin{cases} 3/4 \text{ si } -1 \leq x \leq 0 \\ 1/4 \text{ si } 1 \leq x \leq 2 \\ 0 \quad \text{ailleurs} \end{cases}$$

Quel est le 95^{e} centile de Y?

Question n^{o} 44

La probabilité qu'un article produit par une certaine machine soit conforme aux spécifications techniques est de 0,95, indépendamment d'un article à l'autre. On recueille des articles produits par cette machine jusqu'à ce qu'on ait obtenu *un* article conforme aux spécifications techniques. On répète cette expérience lors de 15 journées consécutives (indépendantes). Soit X le nombre de journées, parmi ces 15 journées, où l'on a dû recueillir au moins 2 articles afin d'obtenir un article conforme aux spécifications techniques.

(a) Quelle est la valeur moyenne de X?

(b) Utiliser une distribution de Poisson pour calculer approximativement la probabilité $P[X = 2 \mid X \geq 1]$.

Question n^{o} 45

Dix échantillons de taille 10 sont tirés au hasard et sans remise de lots identiques contenant 100 articles, dont 2 sont défectueux. On accepte le lot si l'on trouve au plus un article défectueux dans l'échantillon correspondant. Quelle est la probabilité que moins de 9 des 10 lots soient acceptés?

Question n^{o} 46

Le nombre X de particules émises par une certaine source radioactive pendant une heure est une variable aléatoire qui présente une distribution de Poisson de paramètre $\lambda = \ln 5$. De plus, on suppose que les émissions de particules sont indépendantes.

(a) (i) Calculer la probabilité que pendant au moins 30 heures, parmi les 168 heures d'une semaine donnée, aucune particule ne soit émise.

(ii) Utiliser une distribution de Poisson pour calculer approximativement la probabilité en (i).

(b) Calculer la probabilité que la quatrième heure, pendant laquelle aucune particule n'est émise, se produise au cours de la première journée de la semaine considérée en (a).

Question n° 47

La durée X (en heures) des pannes d'électricité majeures, dans une région donnée, présente approximativement une distribution normale de moyenne $\mu = 2$ et d'écart-type $\sigma = 0{,}75$. Trouver la durée x_0 des pannes pour laquelle la probabilité qu'une panne majeure dure au moins 30 minutes de plus que x_0 est égale à 0,06.

Question n° 48

Une variable aléatoire continue X possède la fonction de densité suivante:

$$f_X(x) = \begin{cases} \dfrac{x}{k} e^{-x^2/2k} & \text{si } x > 0 \\ 0 & \text{ailleurs} \end{cases}$$

où $k > 0$ est une constante.

(a) Calculer la moyenne et la variance de X.

(b) Quel est l'effet de la constante k sur la forme de la fonction f_X?

Remarque. On peut calculer (avec un logiciel, si possible) les coefficients β_1 et β_2 pour répondre à cette question.

Question n° 49

Dans une région donnée, la température X (en °C) au cours du mois de septembre présente une distribution normale de paramètres $\mu = 15$ et $\sigma^2 = 25$.

(a) Soit Y la variable aléatoire qui désigne la température, étant donné qu'elle est supérieure à 17 °C; c'est-à-dire que $Y := X \mid \{X > 17\}$. Calculer la fonction de densité de Y.

(b) Calculer (de façon exacte) la probabilité qu'au cours du mois de septembre la température excède 17 °C exactement 10 fois.

Question n° 50

La quantité X de pluie (en millimètres) qui tombe au cours d'une journée, dans une région donnée, est une variable aléatoire telle que

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 3/4 & \text{si } x = 0 \\ 1 - \frac{1}{4} e^{-x^2} & \text{si } x > 0 \end{cases}$$

Calculer (a) la moyenne de X; (b) l'espérance mathématique et la variance de la variable aléatoire $Y := e^{X^2/2}$.

Remarque. La variable X est un exemple de ce que l'on appelle variable aléatoire de *type mixte*, car elle peut prendre la valeur 0 avec une probabilité positive, mais tous les nombres réels positifs ont une probabilité nulle (comme dans le cas d'une variable aléatoire continue). Pour répondre aux questions ci-dessus, il faut se servir à la fois des formules pour les variables aléatoires discrètes et continues (voir page 101).

Question n° 51

Le nombre de coquilles dans une livre de 500 pages présente une distribution de Poisson de paramètre $\lambda = 2$ par page, indépendamment d'une page à l'autre.

(a) Quelle est la probabilité que l'on doive prendre, au hasard et avec remise, plus de 10 pages pour en obtenir 3 qui contiennent chacune 2 coquilles ou plus?

(b) Il y a en fait 20 pages, parmi les 500, qui contiennent, chacune, exactement 5 coquilles.

(i) Si l'on prend 100 pages, au hasard et sans remise, quelle est la probabilité que moins de 5 pages contiennent exactement 5 coquilles?

(ii) Si l'on prend 50 exemplaires identiques de ce livre et si l'on répète l'expérience en (i), quelle est la probabilité que, pour exactement 30 exemplaires du livre, on obtienne moins de 5 pages avec exactement 5 coquilles?

Question n° 52

Un industriel vend un article au prix fixe v. Il rembourse le prix d'achat à tout acheteur qui constate que le poids de l'article est inférieur à un poids donné w_0 et il récupère l'article, dont la valeur de la matière première utilisable est c ($< v$). Le poids W présente approximativement une distribution normale de moyenne μ et de variance σ^2. Un réglage adéquat permet de fixer μ à n'importe quelle valeur désirable, mais il n'est pas possible de fixer la valeur de σ. Le prix de revient R est une fonction du poids de l'article: $R = \alpha + \beta W$, où α et β sont des constantes réelles positives.

(a) Donner une expression pour le bénéfice B en fonction de W.

(b) On peut montrer que le bénéfice moyen, $b(\mu)$, est donné par

$$b(\mu) = v - \alpha - \beta\mu - (v - c)P[W < w_0]$$

Trouver la valeur μ_0 de μ qui maximise $b(\mu)$.

Question n$^{\text{o}}$ 53

Dans une collection de 20 roches, 10 sont de type basalte et 10 sont de type granite. Cinq roches sont prises au hasard et sans remise dans le but d'effectuer des analyses chimiques. Soit X le nombre de roches de type basalte dans l'échantillon.

(a) Donner la distribution de probabilité de X et ses paramètres.

(b) Calculer la probabilité que l'échantillon ne contienne que des roches de même type.

Questions à choix multiple

Question n$^{\text{o}}$ 1

Soit

$$f_X(x) = \begin{cases} 1/2 & \text{si } 0 < x < 1 \\ 1/(2x) & \text{si } 1 \leq x < e \end{cases}$$

Calculer $P[X < 2 \mid X > 1]$.

(a) $\dfrac{\ln 2}{4}$ (b) $\dfrac{\ln 2}{2}$ (c) $\dfrac{2\ln 2}{3}$ (d) $\ln 2$ (e) 1

Question n$^{\text{o}}$ 2

Supposons que $X \sim \text{U}(0,1)$. Calculer $E[(X - E[X])^3]$.

(a) 0 (b) $\frac{1}{4}$ (c) $\frac{1}{3}$ (d) $\frac{1}{2}$ (e) $\frac{2}{3}$

Question n$^{\text{o}}$ 3

Soit $X \sim \text{B}(n = 2, p = 0{,}5)$. Calculer $P[X \geq 1 \mid X \leq 1]$.

(a) 0 (b) $\frac{1}{4}$ (c) $\frac{1}{2}$ (d) $\frac{2}{3}$ (e) 1

Question n$^{\text{o}}$ 4

Calculer $E\left[X^2\right]$ si $p_X(0) = e^{-\lambda}$ et

$$p_X(x) = \frac{e^{-\lambda}}{2}\frac{\lambda^{|x|}}{|x|!} \quad \text{pour } x = \ldots, -2, -1, 1, 2, \ldots$$

où $\lambda > 0$.

(a) λ (b) $\lambda^2 + \lambda$ (c) $\lambda^2 - \lambda$ (d) λ^2 (e) $2\lambda^2$

Question n° 5

Soit

$$f_X(x) = \begin{cases} \frac{1}{2}e^x & \text{si } x < 0 \\ \frac{1}{2}e^{-x} & \text{si } x \geq 0 \end{cases}$$

Calculer la variance de X.

(a) 1 (b) $\frac{3}{2}$ (c) 2 (d) 3 (e) 4

Question n° 6

Supposons que

$$F_X(x) = \begin{cases} 0 & \text{si } x < -1 \\ 1/4 & \text{si } -1 \leq x < 0 \\ 3/4 & \text{si } 0 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

Calculer $p_X(0) + p_X(1)$.

(a) 0 (b) $\frac{1}{4}$ (c) $\frac{1}{2}$ (d) $\frac{3}{4}$ (e) 1

Question n° 7

Soit $f_X(x) = 2xe^{-x^2}$ pour $x > 0$. On pose $Y = \ln X$. Calculer $f_Y(y)$.

(a) $2e^{2y}e^{-e^{2y}}$ (b) $2e^{2y}e^{-y^2}$ (c) $2e^{y}e^{-e^{2y}}$ (d) $2e^{-2y}$ (e) e^{-y}

Question n° 8

Calculer $P[X < 5]$ si $X \sim \mathrm{G}(\alpha = 5, \lambda = 1/2)$.

(a) 0,109 (b) 0,243 (c) 0,5 (d) 0,757 (e) 0,891

Question n° 9

Supposons que X est une variable aléatoire discrète dont l'ensemble des valeurs possibles est $\{0, 1, 2, \ldots\}$. Calculer $P[X = 0]$ si $E\left[t^X\right] = e^{\lambda(t-1)}$, où t est une constante réelle.

(a) 0 (b) 1/4 (c) 1/2 (d) $e^{-2\lambda}$ (e) $e^{-\lambda}$

Question n° 10

Calculer $P\left[X^2 < 4\right]$ si $X \sim \mathrm{Exp}(\lambda = 2)$.

(a) $1 - e^{-4}$ (b) $2\left(1 - e^{-4}\right)$ (c) $\frac{1}{2}e^{-4}$ (d) e^{-4} (e) $2e^{-4}$

Question n° 11

Calculer $P[0 < X < 2]$ si

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{2} & \text{si } x = 0 \\ 1 - \frac{e^{-x}}{2} & \text{si } x > 0 \end{cases}$$

(a) $\dfrac{e^{-1}-e^{-3}}{2}$ (b) $\dfrac{1}{2}-\dfrac{e^{-2}}{2}$ c) $\dfrac{1}{2}$ (d) $1-\dfrac{e^{-2}}{2}$ (e) $1-\dfrac{e^{-3}}{2}$

Remarque. La variable aléatoire X est de *type mixte* (voir page 101 et l'exercice n° 50, page 137). Notons que la fonction F_X est discontinue au point 0.

Question n° 12

On définit $Y=|X|$, où X est une variable aléatoire continue dont la fonction de densité est

$$f_X(x)=\begin{cases}\frac{1}{2} \text{ si } -1\leq x\leq 1\\ 0 \text{ ailleurs}\end{cases}$$

Calculer $f_Y(y)$.

(a) $\dfrac{1}{2}$ si $-1\leq y\leq 1$ (b) $\dfrac{y+1}{2}$ si $-1\leq y\leq 1$ (c) y si $0\leq y\leq 1$

(d) $2y$ si $0\leq y\leq 1$ (e) 1 si $0\leq y\leq 1$

Question n° 13

Soit

x	1	4	9
$p_X(x)$	1/4	1/4	1/2

On a: $E\left[\sqrt{X}\right]=9/4$ et $E[X]=23/4$. Calculer l'écart-type de $\sqrt{X}+4$.

(a) $\frac{\sqrt{11}}{4}$ (b) $\frac{\sqrt{11}}{4}+2$ (c) $\frac{\sqrt{11}}{4}+4$ (d) $\frac{11}{6}$ (e) $\frac{11}{6}+2$

Question n° 14

Supposons que

$$f_X(x)=\begin{cases}1/e \text{ si } 0<x<e\\ 0 \quad \text{ailleurs}\end{cases}$$

Trouver une fonction $g(x)$ telle que si $Y:=g(X)$, alors $f_Y(y)=e^{-y-1}$ si $y>1$.

(a) e^{-x} (b) e^{x-1} (c) e^{x} (d) $-\ln x$ (e) $\ln x$

Question n° 15

Calculer le troisième moment centré de la variable aléatoire discrète X dont la fonction de probabilité est

x	-1	0	1
$p_X(x)$	1/8	1/2	3/8

(a) $-3/32$ (b) 0 (c) 1/64 (d) 3/32 (e) 1/4

Question n° 16

Soit X une variable aléatoire continue définie sur l'intervalle (a, b). Quelle est la fonction de densité de la variable aléatoire $Y := F_X(X)$?

(a) elle n'est pas définie (b) 0 (c) 1 si $0 \leq y \leq 1$ (d) $f_X(y)$
(e) $2y$ si $0 \leq y \leq 1$

Question n° 17

Le taux de suicide dans une certaine ville est de quatre par mois, selon une distribution de Poisson, indépendamment d'un mois à l'autre. Calculer la probabilité qu'il y ait au moins un mois durant l'année pendant lequel il y aura au moins huit suicides.

(a) 0,0520 (b) 0,1263 (c) 0,4731 (d) 0,5269 (e) 0,8737

Question n° 18

Une enquête téléphonique est menée pour connaître l'opinion du public au sujet de la construction d'une centrale nucléaire. Il y a 150.000 abonnés dont le numéro de téléphone est inscrit dans l'annuaire d'une certaine ville et on suppose que 90.000 d'entre eux donneraient une réponse négative si l'on communiquait avec eux. Soit X le nombre de réponses négatives obtenues en 15 appels effectués au hasard. Calculer approximativement $P[X = 9]$ si l'on suppose aussi que l'on ne téléphone pas deux fois au même numéro.

(a) 0,1666 (b) 0,1766 (c) 0,1866 (d) 0,1966 (e) 0,2066

Question n° 19

Un examen à choix multiple comprend 30 questions. Pour chaque question, cinq réponses sont proposées. Toute bonne réponse donne deux points et toute mauvaise réponse fait perdre 1/2 point. Supposons qu'une étudiante a déjà répondu à 20 questions. Ensuite, elle décide de choisir la lettre a pour chacune des 10 questions restantes, sans même les lire. Si les bonnes réponses sont distribuées au hasard parmi les lettres a, b, c, d et e, quelle est la note (sur 60) à laquelle elle peut s'attendre, en supposant qu'elle a 4 chances sur 5 d'avoir réussi n'importe laquelle des 20 premières questions (indépendantes) auxquelles elle a répondu?

(a) 26 (b) 28 (c) 30 (d) 32 (e) 36

Question n° 20

Soit $p_X(x) = (3/4)^{x-1}(1/4)$ pour $x = 1, 2, \ldots$ Calculer l'espérance mathématique de la variable aléatoire discrète X, étant donné que X est supérieure à 2.

(a) 4 (b) 5 (c) 6 (d) 7 (e) 8

4

Vecteurs aléatoires

On peut généraliser la notion de variable aléatoire au cas de deux (ou plusieurs) dimensions. Dans ce manuel, nous n'allons considérer en détail que les *vecteurs aléatoires* de dimension 2. Cependant, l'extension des définitions au cas multi-dimensionnel est immédiate. Nous allons aussi voir dans ce chapitre le théorème le plus important de la théorie des probabilités, soit le *théorème central limite*. Ce chapitre complète la partie du manuel consacrée au calcul des probabilités.

4.1 Vecteurs aléatoires discrets

La **fonction de probabilité conjointe**

$$p_{X,Y}(x_j, y_k) := P\left[\{X = x_j\} \cap \{Y = y_k\}\right] \equiv P[X = x_j, Y = y_k]$$

du couple de variables aléatoires discrètes (X, Y), dont les valeurs possibles sont un ensemble (fini ou infini dénombrable) de couples (x_j, y_k) dans le plan, possède les propriétés suivantes:
(i) $p_{X,Y}(x_j, y_k) \geq 0 \ \forall (x_j, y_k)$;
(ii) $\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} p_{X,Y}(x_j, y_k) = 1$.
La **fonction de répartition conjointe** $F_{X,Y}$ est définie par

$$F_{X,Y}(x, y) = P[\{X \leq x\} \cap \{Y \leq y\}] = \sum_{x_j \leq x} \sum_{y_k \leq y} p_{X,Y}(x_j, y_k) \qquad (4.1)$$

Exemple 4.1.1. Considérons la fonction de probabilité conjointe $p_{X,Y}$ qui est donnée par le tableau suivant:

$y \backslash\, x$	-1	0	1
0	1/16	1/16	1/16
1	1/16	1/16	2/16
2	2/16	1/16	6/16

On peut vérifier que la fonction $p_{X,Y}$ possède les deux propriétés des fonctions de probabilité conjointes énoncées ci-dessus. De plus, étant donné que seuls les couples $(-1,0)$ et $(0,0)$ sont tels que $x_j \leq 0$ *et* $y_k \leq \frac{1}{2}$, on peut écrire que

$$F_{X,Y}\big(0,\tfrac{1}{2}\big) = p_{X,Y}(-1,0) + p_{X,Y}(0,0) = \frac{1}{8}$$

♢

Lorsqu'on fait la somme de la fonction $p_{X,Y}$ sur toutes les valeurs possibles de Y (respectivement X), on obtient ce que l'on appelle **fonction de probabilité marginale** de X (respectivement Y). C'est-à-dire que

$$p_X(x_j) = \sum_{k=1}^{\infty} p_{X,Y}(x_j, y_k) \quad \text{et} \quad p_Y(y_k) = \sum_{j=1}^{\infty} p_{X,Y}(x_j, y_k) \tag{4.2}$$

Exemple 4.1.1 (suite). On trouve que

x	-1	0	1	Σ
$p_X(x)$	1/4	3/16	9/16	1

et

y	0	1	2	Σ
$p_Y(y)$	3/16	1/4	9/16	1

♢

Définition 4.1.1. *Deux variables aléatoires discrètes, X et Y, sont dites* **indépendantes** *si et seulement si*

$$p_{X,Y}(x_j, y_k) = p_X(x_j)p_Y(y_k) \quad \textit{pour tout couple } (x_j, y_k) \tag{4.3}$$

Exemple 4.1.1 (suite). On a: $p_{X,Y}(-1,0) = 1/16$, $p_X(-1) = 1/4$ et $p_Y(0) = 3/16$. Puisque $1/16 \neq (1/4)(3/16)$, X et Y ne sont *pas* des variables aléatoires indépendantes. ♢

Finalement, soit A_X un événement défini en fonction de la variable aléatoire X; par exemple, $A_X = \{X \geq 0\}$. On définit la **fonction de probabilité conditionnelle** de Y, étant donné l'événement A_X, par

$$p_Y(y \mid A_X) \equiv P[Y = y \mid A_X] = \frac{P[\{Y = y\} \cap A_X]}{P[A_X]} \quad \text{si } P[A_X] > 0 \tag{4.4}$$

De même, on définit

$$p_X(x \mid A_Y) \equiv P[X = x \mid A_Y] = \frac{P[\{X = x\} \cap A_Y]}{P[A_Y]} \quad \text{si } P[A_Y] > 0 \tag{4.5}$$

Remarque. Si X et Y sont des variables aléatoires discrètes indépendantes, alors on peut écrire que

$$p_Y(y \mid A_X) \equiv p_Y(y) \quad \text{et} \quad p_X(x \mid A_Y) \equiv p_X(x) \tag{4.6}$$

Exemple 4.1.1 (suite). Soit $A_X = \{X = 1\}$; on a:

$$\begin{aligned} p_Y(y \mid X = 1) &= \frac{P[\{Y = y\} \cap \{X = 1\}]}{P[X = 1]} = \frac{p_{X,Y}(1, y)}{p_X(1)} \\ &= \frac{16}{9} p_{X,Y}(1, y) = \begin{cases} 1/9 \text{ si } y = 0 \\ 2/9 \text{ si } y = 1 \\ 2/3 \text{ si } y = 2 \end{cases} \end{aligned}$$

♢

Exemple 4.1.2. Une boîte contient six transistors, dont un de marque A, deux de marque B et trois de marque C. Deux transistors sont pris au hasard et avec remise. Soit X (respectivement Y) le nombre de transistors de marque A (respectivement B) parmi les deux tirés au hasard.

(a) Calculer la fonction de probabilité conjointe $p_{X,Y}(x, y)$.

(b) Calculer les fonctions de probabilité marginales.

(c) Les variables aléatoires X et Y sont-elles indépendantes? Justifier.

(d) Calculer la probabilité $P[X = Y]$.

Solution. (a) Les valeurs possibles du couple (X, Y) sont: $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$, $(0, 2)$ et $(2, 0)$. Puisque les transistors sont pris *avec* remise, de sorte que les tirages sont indépendants, on obtient le tableau suivant:

$y \backslash\, x$	0	1	2
0	1/4	1/6	1/36
1	1/3	1/9	0
2	1/9	0	0

Par exemple, on a:

$$p_{X,Y}(0,0) = (1/2)(1/2) = 1/4$$

(par indépendance des tirages) et

$$p_{X,Y}(1,0) \overset{\text{inc.}}{=} P[A_1 \cap C_2] + P[C_1 \cap A_2] \overset{\text{ind.,sym.}}{=} 2(1/6)(1/2) = 1/6$$

où A_k: un transistor de marque A est obtenu lors du k^{e} tirage, etc. On peut vérifier que la somme de toutes les fractions dans le tableau égale 1.

Remarque. En fait, les variables aléatoires X et Y présentent des distributions binomiales de paramètres $n = 2$ et $p = 1/6$, et de paramètres $n = 2$ et $p = 1/3$, respectivement.

(b) À partir du tableau en (a), on trouve que

x	0	1	2	Σ
$p_X(x)$	25/36	5/18	1/36	1

et

y	0	1	2	Σ
$p_Y(y)$	4/9	4/9	1/9	1

(c) Les variables aléatoires X et Y ne sont *pas* indépendantes, car, par exemple,

$$p_{X,Y}(2,2) = 0 \neq p_X(2)p_Y(2) = (1/36)(1/9)$$

Remarque. Les variables aléatoires X et Y ne pouvaient pas être indépendantes, car la relation $0 \leq X + Y \leq 2$ doit être vérifiée.

(d) On calcule

$$P[X = Y] = p_{X,Y}(0,0) + p_{X,Y}(1,1) + p_{X,Y}(2,2) = \frac{1}{4} + \frac{1}{9} + 0 = \frac{13}{36}$$

♢

4.2 Vecteurs aléatoires continus

Soit X et Y deux variables aléatoires continues; la généralisation de la notion de fonction de densité au cas de dimension 2 est la **fonction de densité conjointe** $f_{X,Y}$ du couple (X, Y) qui possède les propriétés suivantes:
(i) $f_{X,Y}(x, y) \geq 0$ pour tout couple (x, y);
(ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dxdy = 1$.

Remarque. En n dimensions, un vecteur aléatoire continu possède une fonction de densité conjointe définie sur $\mathbb{R}^n$ (ou un sous-ensemble infini non dénombrable de $\mathbb{R}^n$). Cette fonction doit être non négative et son intégrale sur $\mathbb{R}^n$ doit égaler 1.

La **fonction de répartition conjointe** est définie par

$$F_{X,Y}(x, y) = P[X \leq x, Y \leq y] = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, dudv \tag{4.7}$$

Exemple 4.2.1. Considérons la fonction $f_{X,Y}$ définie par

$$f_{X,Y}(x, y) = c\, xye^{-x^2-y^2} \quad \text{pour } x \geq 0, y \geq 0$$

où $c > 0$ est une constante. On a: (i) $f_{X,Y}(x, y) \geq 0$ pour tout couple (x, y) avec $x \geq 0$ et $y \geq 0$ ($f_{X,Y}(x, y) = 0$ ailleurs) et (ii)

$$\int_0^{\infty} \int_0^{\infty} c\, xye^{-x^2-y^2} dxdy = c \int_0^{\infty} xe^{-x^2} dx \int_0^{\infty} ye^{-y^2} dy$$
$$= c \left[-\frac{e^{-x^2}}{2} \Big|_0^{\infty} \right] \left[-\frac{e^{-y^2}}{2} \Big|_0^{\infty} \right] = c(1/2)(1/2) = c/4$$

Donc, cette fonction est une fonction de densité conjointe valable si et seulement si la constante c égale 4.

La fonction de répartition conjointe du couple (X, Y) est donnée par

$$\begin{aligned} F_{X,Y}(x, y) &= \int_0^{y} \int_0^{x} 4\, uve^{-u^2-v^2} dudv \\ &= \int_0^{x} 2ue^{-u^2} du \int_0^{y} 2ve^{-v^2} dv \\ &= \left[-e^{-u^2} \Big|_0^{x} \right] \left[-e^{-v^2} \Big|_0^{y} \right] = (1 - e^{-x^2})(1 - e^{-y^2}) \end{aligned}$$

pour $x \geq 0$ *et* $y \geq 0$.

Remarque. On a: $F_{X,Y}(x,y) = 0$ si $x < 0$ ou $y < 0$. ♢

Les **fonctions de densité marginales** de X et Y sont définies par

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \quad \text{et} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx \tag{4.8}$$

Remarque. On peut généraliser facilement les définitions ci-dessus au cas de trois variables aléatoires ou plus. Ainsi, la fonction de densité conjointe du vecteur aléatoire (X,Y,Z) est une fonction $f_{X,Y,Z}(x,y,z)$ non négative et telle que

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,dxdydz = 1$$

De plus, la fonction de densité conjointe du vecteur aléatoire (X,Y) est obtenue comme suit:

$$f_{X,Y}(x,y) = \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,dz$$

Finalement, la fonction de densité marginale de la variable aléatoire X est donnée par

$$f_X(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)\,dydz$$

Définition 4.2.1. *Les variables aléatoires continues X et Y sont dites* **indépendantes** *si et seulement si*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \textit{pour tout couple } (x,y) \tag{4.9}$$

Exemple 4.2.1 (suite). On a:

$$\begin{aligned} f_X(x) &= \int_0^{\infty} 4\,xye^{-x^2-y^2}dy = 2xe^{-x^2}\int_0^{\infty} 2ye^{-y^2}dy \\ &= 2xe^{-x^2}\left[-e^{-y^2}\Big|_0^{\infty}\right] = 2xe^{-x^2} \quad \text{pour } x \geq 0 \end{aligned}$$

Alors, par symétrie, on peut écrire que

$$f_Y(y) = 2ye^{-y^2} \quad \text{pour } y \geq 0$$

De plus, puisque

$$f_X(x)f_Y(y) = 4xye^{-x^2-y^2} = f_{X,Y}(x,y)$$

pour tout couple (x, y), les variables aléatoires X et Y sont indépendantes. $\diamondsuit$

Finalement, soit A_Y un événement qui ne dépend que de Y; par exemple, $A_Y = \{0 \leq Y \leq 1\}$. La **fonction de densité conditionnelle** de X, étant donné que A_Y s'est produit, est donnée par

$$f_X(x \mid A_Y) = \frac{\int_{A_Y} f_{X,Y}(x,y)dy}{P[A_Y]} \quad \text{si } P[A_Y] > 0 \tag{4.10}$$

Si A_Y est un événement de la forme $\{Y = y\}$, on peut montrer que

$$f_X(x \mid Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{si } f_Y(y) > 0 \tag{4.11}$$

C'est-à-dire que la fonction de densité conditionnelle, $f_X(x \mid Y = y)$, est obtenue en divisant la fonction de densité conjointe de (X, Y), évaluée au point (x, y), par la fonction de densité marginale de Y évaluée au point y.

Remarques.
(i) Si X et Y sont deux variables aléatoires continues indépendantes, alors on a:

$$f_X(x \mid A_Y) \equiv f_X(x) \quad \text{et} \quad f_Y(y \mid A_X) \equiv f_Y(y) \tag{4.12}$$

(ii) En général, si X est une variable aléatoire continue, alors on peut écrire que

$$P[Y \in A_Y] = \int_{-\infty}^{\infty} P[Y \in A_Y \mid X = x] f_X(x)dx \tag{4.13}$$

où A_Y est un événement qui n'implique que la variable aléatoire Y. Dans le cas où X est de type discret, on a:

$$P[Y \in A_Y] = \sum_{k=1}^{\infty} P[Y \in A_Y \mid X = x_k] p_X(x_k) \tag{4.14}$$

Ces formules sont des extensions de la *règle de la probabilité totale* du chapitre 2. On a aussi, par exemple:

$$P[Y > X] = \int_{-\infty}^{\infty} P[Y > X \mid X = x] f_X(x) dx$$
$$= \int_{-\infty}^{\infty} P[Y > x \mid X = x] f_X(x) dx \qquad (4.15)$$

etc.

Définition 4.2.2. *L'***espérance conditionnelle** *de la variable aléatoire Y, étant donné que $X = x$, est définie par*

$$E[Y \mid X = x] = \begin{cases} \sum_{j=1}^{\infty} y_j p_Y(y_j \mid X = x) \text{ si } (X,Y) \text{ est discret} \\ \\ \int_{-\infty}^{\infty} y f_Y(y \mid X = x) dy \text{ si } (X,Y) \text{ est continu} \end{cases}$$

Remarques.
(i) On peut montrer que

$$E[Y] = E[E[Y \mid X]] := \begin{cases} \sum_{k=1}^{\infty} E[Y \mid X = x_k] p_X(x_k) \text{ si } X \text{ est discrète} \\ \\ \int_{-\infty}^{\infty} E[Y \mid X = x] f_X(x) dx \text{ si } X \text{ est continue} \end{cases}$$

(ii) En général,

$$E[g(Y)] = E[E[g(Y) \mid X]]$$

pour n'importe quelle fonction $g(\cdot)$. Il s'ensuit que

$$\text{VAR}[Y] = E[E[Y^2 \mid X]] - \{E[E[Y \mid X]]\}^2$$

Exemple 4.2.2. Soit

$$f_{X,Y}(x,y) = \begin{cases} k(x^2 + y^2) \text{ pour } 0 \le x \le a,\ 0 \le y \le b \\ 0 \qquad \text{ailleurs} \end{cases}$$

Calculer

(a) la constante k;

(b) les fonctions de densité marginales de X et Y; X et Y sont-elles indépendantes?

(c) les fonctions de densité conditionnelles $f_X(x \mid Y = y)$, $f_Y(y \mid X = x)$ et $f_Y(y \mid X < \frac{a}{2})$;

(d) la fonction de répartition $F_{X,Y}(x, y)$.

Solution. (a) On a:

$$\begin{aligned} 1 &= \int_0^b \int_0^a k(x^2 + y^2)dxdy = k \int_0^b \left(\frac{x^3}{3} + xy^2 \right) \Big|_0^a dy \\ &= k \int_0^b \left(\frac{a^3}{3} + ay^2 \right) dy = k \left[\frac{a^3 y}{3} + \frac{ay^3}{3} \right] \Big|_0^b \end{aligned}$$

Donc, il faut que

$$k = \left[\frac{a^3 b}{3} + \frac{ab^3}{3} \right]^{-1} = \frac{3}{(ab)(a^2 + b^2)}$$

(b) On peut écrire que

$$\begin{aligned} f_X(x) &= \int_0^b k(x^2 + y^2)dy = k \left\{ yx^2 + \frac{y^3}{3} \Big|_0^b \right\} \\ &= k \left(bx^2 + \frac{b^3}{3} \right) \quad \text{pour } 0 \leq x \leq a \end{aligned}$$

où k a été calculé en (a). De même, on trouve que

$$f_Y(y) = k \left(ay^2 + \frac{a^3}{3} \right) \quad \text{pour } 0 \leq y \leq b$$

Maintenant, on a:

$$f_X(0) f_Y(0) = k \left(\frac{b^3}{3} \right) k \left(\frac{a^3}{3} \right) = k^2 \frac{a^3 b^3}{9} \neq 0 = f_{X,Y}(0, 0)$$

Alors X et Y ne sont *pas* des variables aléatoires indépendantes.

Remarque. Lorsque la fonction de densité conjointe, $f_{X,Y}(x, y)$, est une constante c multipliée par une somme ou une différence, comme $x + y$, $x^2 - y^2$, etc., les variables aléatoires X et Y ne peuvent pas être indépendantes. En effet, il est impossible d'écrire, par exemple, que

$$c(x + y) = f(x)g(y)$$

où $f(x)$ ne dépend que de x et $g(y)$ ne dépend que de y.

(c) On calcule

$$f_X(x \mid Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \overset{\text{(b)}}{=} \frac{k(x^2+y^2)}{k\left(\frac{a^3}{3} + ay^2\right)} = \frac{3(x^2+y^2)}{a(a^2+3y^2)}$$

pour $0 \le x \le a$ et $0 \le y \le b$. De même, on trouve que

$$f_Y(y \mid X = x) = \frac{3(x^2+y^2)}{b(3x^2+b^2)}$$

pour $0 \le x \le a$ et $0 \le y \le b$. Finalement, on a:

$$\begin{aligned} f_Y\left(y \mid X < \tfrac{a}{2}\right) &= \frac{\int_0^{a/2} k(x^2+y^2)dx}{\int_0^{a/2} k\left(bx^2 + \frac{b^3}{3}\right)dx} = \frac{k\left\{\frac{x^3}{3} + xy^2\Big|_0^{a/2}\right\}}{k\left\{b\frac{x^3}{3} + x\frac{b^3}{3}\Big|_0^{a/2}\right\}} \\ &= \frac{\frac{a^3}{24} + \frac{ay^2}{2}}{\frac{ba^3}{24} + \frac{ab^3}{6}} = \frac{a^2+12y^2}{ba^2+4b^3} \quad \text{pour } 0 \le y \le b \end{aligned}$$

(d) Par définition,

$$\begin{aligned} F_{X,Y}(x,y) &= \int_0^y \int_0^x k(u^2+v^2)dudv = \int_0^y k\left[\frac{u^3}{3} + uv^2\right]\Big|_0^x dv \\ &= k\int_0^y \left(\frac{x^3}{3} + xv^2\right)dv = k\left\{\frac{x^3y}{3} + \frac{xy^3}{3}\right\} = \frac{xy(x^2+y^2)}{ab(a^2+b^2)} \end{aligned}$$

pour $0 \le x \le a$ *et* $0 \le y \le b$. De là, on déduit que (voir la figure 4.1)

$$F_{X,Y}(x,y) = \begin{cases} 0 & \text{si } x < 0 \text{ ou } y < 0 \\ \dfrac{xy(x^2+y^2)}{ab(a^2+b^2)} & \text{pour } 0 \le x \le a,\ 0 \le y \le b \\ \dfrac{xb(x^2+b^2)}{ab(a^2+b^2)} & \text{pour } 0 \le x \le a,\ y > b \\ \dfrac{ay(a^2+y^2)}{ab(a^2+b^2)} & \text{pour } x > a,\ 0 \le y \le b \\ 1 & \text{si } x > a \text{ et } y > b \end{cases}$$

♢

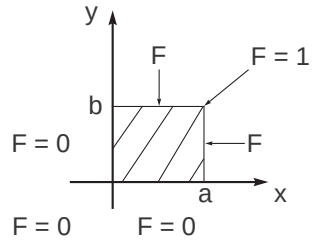

Fig. 4.1. Fonction de répartition conjointe dans l'exemple 4.2.2

Remarque. Comme dans le cas unidimensionnel, on a:

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = f_{X,Y}(x,y) \tag{4.16}$$

pour tout couple (x, y) où la fonction $F_{X,Y}(x, y)$ est dérivable.

Exemple 4.2.3. Supposons que $X \sim \text{Exp}(\lambda_1)$ et $Y \sim \text{Exp}(\lambda_2)$ sont des variables aléatoires indépendantes. En utilisant la formule (4.15), on peut écrire que

$$\begin{aligned} P[Y > X] &= \int_0^\infty P[Y > X \mid X = x] f_X(x)\, dx \\ &\overset{\text{ind.}}{=} \int_0^\infty P[Y > x] \lambda_1 e^{-\lambda_1 x}\, dx = \int_0^\infty e^{-\lambda_2 x} \lambda_1 e^{-\lambda_1 x}\, dx \\ &= \int_0^\infty \lambda_1 e^{-(\lambda_1+\lambda_2)x}\, dx = \frac{\lambda_1}{\lambda_1 + \lambda_2} \end{aligned} \tag{4.17}$$

Remarque. Notons que si $\lambda_1 = \lambda_2$, alors $P[X < Y] = 1/2$, ce qu'on pouvait en fait affirmer directement par symétrie (et par continuité de la distribution expoentielle).

4.3 Fonctions de vecteurs aléatoires

Au chapitre 3, nous avons vu que toute fonction à valeurs réelles d'une variable aléatoire est elle-même une variable aléatoire. De même, toute fonction à valeurs réelles d'un vecteur aléatoire est une variable aléatoire. De façon plus générale, si l'on applique n fonctions à valeurs réelles à une variable ou à un vecteur aléatoire, alors on obtient un nouveau vecteur aléatoire de dimension n.

Les transformations qui nous intéressent le plus sont la somme, la différence, le produit et le quotient des variables aléatoires.

En général, il faut savoir calculer la fonction de probabilité ou la fonction de densité de la nouvelle variable ou du nouveau vecteur aléatoire. Dans le cadre de ce livre, nous allons traiter le cas où l'on applique une seule fonction g à un vecteur aléatoire (X, Y) de dimension $n = 2$. Nous donnerons aussi des résultats importants obtenus lorsque la fonction g est la somme (ou bien une *combinaison linéaire*) de n variables aléatoires indépendantes.

Parfois, on désire obtenir uniquement la moyenne, par exemple, de la nouvelle variable aléatoire. Dans ce cas, il n'est pas nécessaire de calculer d'abord la fonction de probabilité ou la fonction de densité de $g(X, Y)$.

4.3.1 Cas discret

Dans le cas particulier où le nombre de valeurs possibles du couple de variables aléatoires (X, Y) est *fini*, il suffit d'appliquer la transformation g à chacun des couples possibles et d'additionner les probabilités des couples (x, y) qui sont transformés en la même valeur de $g(x, y)$.

Exemple 4.3.1. Reprenons la fonction de probabilité conjointe de l'exemple 4.1.1:

$y \backslash\, x$	-1	0	1
0	1/16	1/16	1/16
1	1/16	1/16	2/16
2	2/16	1/16	6/16

Soit $Z = XY$. La variable aléatoire Z peut prendre cinq valeurs différentes: -2, -1, 0, 1 et 2. Le couple $(-1, 2)$ correspond à $z = -2$, $(-1, 1)$ correspond à $z = -1$, $(1, 1)$ est transformé en $z = 1$, $(1, 2)$ devient $z = 2$, et tous les autres couples donnent $z = 0$. À partir du tableau ci-dessus, on obtient que

z	-2	-1	0	1	2	Σ
$p_Z(z)$	2/16	1/16	5/16	2/16	6/16	1

Il s'ensuit que la moyenne de Z est donnée par

$$E[Z] = (-2)\frac{2}{16} + (-1)\frac{1}{16} + 0 + (1)\frac{2}{16} + (2)\frac{6}{16} = \frac{9}{16}$$

♢

Comme nous l'avons mentionné ci-dessus, si tout ce qui nous intéresse est d'obtenir la moyenne de la nouvelle variable aléatoire Z, alors il n'est pas nécessaire de calculer la fonction $p_Z(z)$. Il suffit d'utiliser la formule (4.31) de la section 4.4:

$$E[g(X,Y)] = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} g(x_k, y_j) p_{X,Y}(x_k, y_j)$$

Ici, on obtient que $E[XY] = 9/16$ (voir l'exemple 4.4.1), ce qui concorde avec le résultat obtenu ci-dessus avec $Z = XY$.

Exemple 4.3.2. Supposons qu'on lance deux tétraèdres distincts et bien équilibrés, dont les faces sont numérotées 1, 2, 3 et 4. Soit X_1 (respectivement X_2) le numéro de la face sur laquelle repose le premier (respectivement deuxième) tétraèdre, et soit Y le *maximum* entre X_1 et X_2. Quelle est la fonction de probabilité de la variable aléatoire Y?

Solution. Les valeurs possibles de Y sont 1, 2, 3 et 4. Soit A_k (respectivement B_k): la variable aléatoire X_1 (respectivement X_2) prend la valeur k, pour $k = 1, \ldots, 4$. Par indépendance des événements A_j et B_k pour tout j et k, on a:

$$p_Y(1) = P[A_1 \cap B_1] = P[A_1]P[B_1] = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

De même, par indépendance et *incompatibilité*, on peut écrire que

$$\begin{aligned} p_Y(2) &= P[(A_1 \cap B_2) \cup (A_2 \cap B_1) \cup (A_2 \cap B_2)] \\ &= P[A_1]P[B_2] + P[A_2]P[B_1] + P[A_2]P[B_2] \\ &= \frac{1}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{4} + \frac{1}{4} \times \frac{1}{4} = \frac{3}{16} \end{aligned}$$

Ensuite, en utilisant l'*équiprobabilité* des événements (et des intersections), on obtient que

$$\begin{aligned} p_Y(3) &= P[(A_1 \cap B_3) \cup (A_2 \cap B_3) \cup (A_3 \cap B_3) \cup (A_3 \cap B_2) \cup (A_3 \cap B_1)] \\ &= 5 \times \frac{1}{4} \times \frac{1}{4} = \frac{5}{16} \end{aligned}$$

Finalement, puisque l'on doit avoir: $\sum_{y=1}^{4} p_Y(y) = 1$, on obtient le tableau suivant:

y	1	2	3	4
$p_Y(y)$	1/16	3/16	5/16	7/16

Il s'ensuit que

y	1	2	3	4
$F_Y(y)$	1/16	1/4	9/16	1

♢

Lorsque le nombre de valeurs possibles du couple (X, Y) est *infini dénombrable*, le travail nécessaire pour obtenir la fonction de probabilité de la variable aléatoire $Z := g(X, Y)$ est généralement beaucoup plus difficile. En effet, il peut y avoir une infinité de couples (x, y) qui correspondent au même $z = g(x, y)$, et il peut aussi y avoir une infinité de z différents. Cependant, dans le cas où le nombre de valeurs possibles de Z est fini, on peut parfois calculer $p_Z(z)$ assez facilement.

Exemple 4.3.3. Supposons que la fonction de probabilité conjointe du vecteur aléatoire (X, Y) est donnée par la formule

$$p_{X,Y}(x, y) = \frac{e^{-2}}{x!y!} \quad \text{pour } x = 0, 1, \ldots ; y = 0, 1, \ldots$$

Notons que X et Y sont en fait deux variables aléatoires indépendantes qui présentent toutes les deux une distribution de Poisson de paramètre $\lambda = 1$. Soit

$$Z = g(X, Y) := \begin{cases} 1 \text{ si } X = Y \\ 0 \text{ si } X \neq Y \end{cases}$$

Dans ce cas, Z présente une distribution de Bernoulli de paramètre p, où

$$p := P[X = Y] = \sum_{x=0}^{\infty} \frac{e^{-2}}{(x!)^2}$$

On trouve, avec un logiciel mathématique par exemple, que la série infinie ci-dessus converge vers $e^{-2} \cdot I_0(2) \simeq 0{,}3085$, où $I_\nu(x)$ est une *fonction de Bessel*. On pouvait en fait obtenir une très bonne approximation du résultat exact en faisant la somme des cinq premiers termes de la série puisque

$$\sum_{x=0}^{4} \frac{e^{-2}}{(x!)^2} \simeq 0{,}3085$$

aussi. ♢

4.3.2 Cas continu

Supposons que l'on applique une transformation $Z := g_1(X,Y)$ au vecteur aléatoire continu (X,Y) et que l'on désire obtenir la fonction de densité de Z. Nous n'allons considérer que le cas où il est possible de définir une *variable auxiliaire* $W = g_2(x,y)$ telle que le système

$$\begin{bmatrix} z = g_1(x,y) \\ w = g_2(x,y) \end{bmatrix} \tag{4.18}$$

possède la solution unique $x = h_1(z,w)$, $y = h_2(z,w)$. On peut alors démontrer la proposition suivante.

Proposition 4.3.1. *Soit (X,Y) un vecteur aléatoire continu. On définit $Z = g_1(X,Y)$ et $W = g_2(X,Y)$. Supposons que les fonctions $x = h_1(z,w)$ et $y = h_2(z,w)$ possèdent des dérivées partielles (par rapport à z et w) continues $\forall (z,w)$ et que le* jacobien *de la transformation:*

$$J(z,w) := \begin{vmatrix} \partial h_1/\partial z & \partial h_1/\partial w \\ \partial h_2/\partial z & \partial h_2/\partial w \end{vmatrix} \tag{4.19}$$

n'est pas identique à zéro. Alors on peut écrire que

$$f_{Z,W}(z,w) = f_{X,Y}(h_1(z,w), h_2(z,w))|J(z,w)| \tag{4.20}$$

Il s'ensuit que

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(h_1(z,w), h_2(z,w))|J(z,w)|\, dw \tag{4.21}$$

Remarques.
(i) On choisit généralement une variable auxiliaire W qui est très simple, comme $W = X$.

(ii) Dans le cas particulier où $g_1(x,y)$ est une transformation linéaire de x et y, il suffit de prendre une autre transformation linéaire de x et y pour que les dérivées partielles des fonctions h_1 et h_2 soient continues. En effet, ces dérivées partielles sont alors des constantes; par conséquent, elles sont continues pour tout couple (z,w).

Exemple 4.3.4. Soit $X \sim$ U(0,1) et $Y \sim$ U(0,1) deux variables aléatoires indépendantes, et soit $Z = X + Y$. Pour obtenir la fonction de densité de Z, on définit la variable auxiliaire $W = X$. Alors le système

$$\begin{bmatrix} z = x + y \\ w = \quad x \end{bmatrix}$$

possède la solution unique $x = w$, $y = z - w$. De plus, les dérivées partielles des fonctions $h_1(z, w) = w$ et $h_2(z, w) = z - w$ sont continues $\forall(z, w)$ et le jacobien

$$J(z, w) = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1 \neq 0 \quad \forall(z, w)$$

Par conséquent, on peut écrire que

$$\begin{aligned} & f_{Z,W}(z, w) = f_{X,Y}(w, z - w)|-1| \overset{\text{ind.}}{=} f_X(w) f_Y(z - w) \\ \Longrightarrow \quad & f_{Z,W}(z, w) = 1 \cdot 1 \quad \text{si } 0 < w < 1 \text{ et } 0 < z - w < 1 \end{aligned}$$

Puisque $0 < z < 2$, les valeurs possibles de w sont $w \in (0, z)$ si $0 < z < 1$, et $w \in (z - 1, 1)$ si $1 \leq z < 2$.

Finalement, on a (voir la figure 4.2):

$$f_Z(z) = \begin{cases} \displaystyle\int_0^z 1 \, dw = z & \text{si } 0 < z < 1 \\ \displaystyle\int_{z-1}^1 1 \, dw = 2 - z & \text{si } 1 \leq z < 2 \end{cases}$$

♢

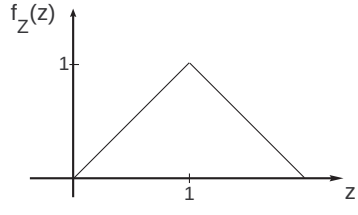

Fig. 4.2. Fonction de densité dans l'exemple 4.3.4

4.3.3 Convolutions

Soit X une variable aléatoire discrète dont les valeurs possibles sont x_1, x_2, ... La **convolution** de X avec elle-même est obtenue en appliquant la transformation qui nous intéresse, soit la somme, la différence, le produit, etc., aux couples $(x_1, x_1), (x_1, x_2), \ldots, (x_2, x_1), (x_2, x_2), \ldots$ Donc, si X peut prendre n valeurs

différentes, alors on doit appliquer la transformation à $n \times n = n^2$ couples. On écrira $X \otimes \mathrm{X}$ pour symboliser le *produit de convolution* de X avec elle-même, $X \oplus X$ pour la *somme de convolution*, etc. Remarquons qu'obtenir la distribution de $X \otimes X$, par exemple, revient à trouver la distribution du produit $X_1 X_2$, où X_1 et X_2 sont deux variables aléatoires indépendantes qui possèdent la même distribution que X.

Exemple 4.3.5. Considérons la fonction de probabilité de la variable aléatoire X de l'exemple 3.4.1:

x	-1	0	1
$p_X(x)$	$1/4$	$1/4$	$1/2$

Supposons que l'on désire obtenir la distribution du produit de convolution de X avec elle-même. On trouve que les résultats possibles de cette convolution sont: $-1, 0$ et 1. Les couples $(-1, 1)$ et $(1, -1)$ correspondent à -1, les couples $(-1, -1)$ et $(1, 1)$ à 1, et les cinq autres couples à 0. Alors, si l'on définit $Y = X \otimes X$, on déduit du tableau ci-dessus que

y	-1	0	1
$p_Y(y)$	$1/4$	$7/16$	$5/16$

car

$$\begin{aligned} P[Y = -1] &= P[X = -1]P[X = 1] + P[X = 1]P[X = -1] \\ &= 2 \times (1/4)(1/2) = 1/4 \end{aligned}$$

etc.

Remarques.

(i) Notons que le résultat obtenu est complètement différent de la fonction de probabilité de $Z := X^2$ calculée dans l'exemple 3.4.1:

z	0	1
$p_Z(z)$	$1/4$	$3/4$

(ii) Si l'on calcule la *différence de convolution* de X avec elle-même, on trouve que la distribution de $D := X \ominus X$ est donnée par

d	-2	-1	0	1	2
$p_D(d)$	$1/8$	$3/16$	$3/8$	$3/16$	$1/8$

$\diamondsuit$

En général, il est difficile de calculer la distribution de la convolution d'une variable aléatoire discrète X avec elle-même k fois, où k est quelconque, surtout si le nombre de valeurs que X peut prendre est infini (dénombrable). Cependant, dans quelques cas particuliers on peut obtenir une formule générale, valable pour tout k entier. En fait, on peut démontrer certains résultats au sujet de la somme de variables aléatoires discrètes indépendantes, mais ne possédant pas nécessairement les mêmes paramètres. Les résultats de ce type les plus importants sont les suivants, où $X_1, \ldots, X_n$ sont n variables aléatoires indépendantes.
(1) Si X_i présente une distribution de Bernoulli de paramètre p pour tout i, alors on a:

$$\sum_{i=1}^{n} X_i \sim \mathrm{B}(n,p) \quad \text{pour } n = 1, 2, \ldots \tag{4.22}$$

De façon plus générale, si $X_i \sim \mathrm{B}(m_i, p)$ pour $i = 1, \ldots, n$, alors on trouve que

$$\sum_{i=1}^{n} X_i \sim \mathrm{B}\left(\sum_{i=1}^{n} m_i, p\right) \tag{4.23}$$

(2) Si $X_i \sim \mathrm{Poi}(\lambda_i)$ pour $i = 1, \ldots, n$, alors on peut écrire que

$$\sum_{i=1}^{n} X_i \sim \mathrm{Poi}\left(\sum_{i=1}^{n} \lambda_i\right) \tag{4.24}$$

(3) Si $X_i \sim \mathrm{Géom}(p)$ pour $i = 1, \ldots, n$, alors on a:

$$\sum_{i=1}^{n} X_i \sim \mathrm{BN}(n,p) \tag{4.25}$$

Supposons maintenant que X et Y sont deux variables aléatoires continues et indépendantes. Soit $Z = X + Y$. On peut montrer que la fonction de densité de Z est obtenue en faisant le produit de convolution de la fonction de densité de X avec celle de Y. C'est-à-dire que l'on a:

$$f_Z(z) = f_X(x) * f_Y(y) = \int_{-\infty}^{\infty} f_X(u) f_Y(z-u)\, du \tag{4.26}$$

On pourrait utiliser cette formule pour obtenir la fonction de densité de Z dans l'exemple 4.3.4.

Comme dans le cas discret, on peut démontrer certains résultats au sujet de la somme, ou parfois de *combinaisons linéaires*, de variables aléatoires X_i *indépendantes*. On trouve, en particulier, que

(1) Si $X_i \sim \text{Exp}(\lambda)$ pour $i = 1, \ldots, n$, alors

$$\sum_{i=1}^{n} X_i \sim \text{G}(n, \lambda) \tag{4.27}$$

(2) Si $X_i \sim \text{G}(\alpha_i, \lambda)$ pour $i = 1, \ldots, n$, alors

$$\sum_{i=1}^{n} X_i \sim \text{G}\left(\sum_{i=1}^{n} \alpha_i, \lambda\right) \tag{4.28}$$

(3) Si $X_i \sim \text{N}(\mu_i, \sigma_i^2)$ pour $i = 1, \ldots, n$, alors

$$\sum_{i=1}^{n} a_i X_i \sim \text{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) \tag{4.29}$$

où les a_i sont des constantes réelles pour tout i.

Remarques.

(i) La meilleure façon de démontrer ces résultats est de se servir des *fonctions caractéristiques*, ou encore des *fonctions génératrices des moments*, qui sont en fait des espérances mathématiques particulières. Ainsi, on définit la fonction caractéristique de la variable aléatoire X par $E[e^{j\omega X}]$, où $j := \sqrt{-1}$.

(ii) Un résultat du même genre, mais pour le *produit* de variables aléatoires indépendantes $X_1, \ldots, X_n$, est le suivant: si $X_i \sim \text{LN}(\mu_i, \sigma_i^2)$, alors

$$\prod_{i=1}^{n} X_i^{a_i} \sim \text{LN}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) \tag{4.30}$$

4.4 Covariance et coefficient de corrélation

Définition 4.4.1. *Soit (X, Y) un couple de variables aléatoires; on définit l'***espérance mathématique** *de la fonction $g(X, Y)$ par*

$$E[g(X,Y)] = \begin{cases} \displaystyle\sum_{k=1}^{\infty}\sum_{j=1}^{\infty} g(x_k,y_j)p_{X,Y}(x_k,y_j) \quad \textit{(cas discret)} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)\,dxdy \textit{ (cas continu)} \end{cases} \tag{4.31}$$

Cas particuliers.
(i) Si $g(X,Y) = X$, alors on a: $E[g(X,Y)] = E[X] \equiv \mu_X$.
(ii) Si $g(X,Y) = XY$, on obtient alors la formule qui permet de calculer l'espérance d'un produit, dont on se sert dans le calcul de la *covariance* et du *coefficient de corrélation.*
(iii) Si la fonction g est une *combinaison linéaire* des variables aléatoires X et Y, c'est-à-dire si l'on a:

$$g(X,Y) = aX + bY + c \tag{4.32}$$

où a, b et c sont des constantes réelles, alors on montre facilement que

$$E[g(X,Y)] = aE[X] + bE[Y] + c \tag{4.33}$$

si les espérances mathématiques existent. Cette formule peut être généralisée au cas de combinaisons linéaires de n variables aléatoires $X_1,\ldots,X_n$. De plus, si $Y = h(X)$, par exemple $Y = X^2$, alors la formule nous permet d'écrire que

$$E[aX + bX^2] = aE[X] + bE[X^2] \tag{4.34}$$

comme nous l'avons fait au chapitre 3.

Définition 4.4.2. *La* **covariance** *de X et Y est définie par*

$$\mathrm{COV}[X,Y] \equiv \sigma_{X,Y} = E[(X-\mu_X)(Y-\mu_Y)] \tag{4.35}$$

Remarques.
(i) On peut montrer que

$$\mathrm{COV}[X,Y] = E[XY] - E[X]E[Y] \tag{4.36}$$

(ii) Si X et Y sont deux variables aléatoires *indépendantes* et si l'on peut écrire que $g(X,Y) = g_1(X)g_2(Y)$, alors on a: $E[g(X,Y)] = E[g_1(X)]E[g_2(Y)]$. Il s'ensuit que si X et Y sont indépendantes, alors

$$\mathrm{COV}[X,Y] \overset{\text{ind.}}{=} E[X]E[Y] - E[X]E[Y] = 0 \tag{4.37}$$

Si la covariance des variables aléatoires X et Y est nulle, elles ne sont *pas nécessairement* indépendantes. Toutefois, on peut montrer que, si X et Y sont deux variables aléatoires qui présentent une distribution *normale*, alors X et Y sont indépendantes *si et seulement si* $\mathrm{COV}[X,Y] = 0$.

(iii) On a: $\mathrm{COV}[X,X] = E[X^2] - (E[X])^2 = \mathrm{VAR}[X]$. Donc, la variance est un cas particulier de la covariance; cependant, contrairement à la variance, la covariance peut être négative.

(iv) Si $g(X,Y)$ est une combinaison linéaire de X et Y, alors on trouve que

$$\mathrm{VAR}[g(X,Y)] \equiv \mathrm{VAR}[aX + bY + c] = a^2\mathrm{VAR}[X] + b^2\mathrm{VAR}[Y] + 2ab\mathrm{COV}[X,Y] \tag{4.38}$$

Notons que la constante c n'influence pas la variance de $g(X,Y)$. De plus, si X et Y sont des variables aléatoires indépendantes, alors on a:

$$\mathrm{VAR}[aX + bY + c] \overset{\text{ind.}}{=} a^2\mathrm{VAR}[X] + b^2\mathrm{VAR}[Y] \tag{4.39}$$

Finalement, la formule (4.38) se généralise au cas où l'on a une combinaison linéaire de n variables aléatoires (indépendantes ou non).

Définition 4.4.3. *Le* **coefficient de corrélation** *de X et Y est donné par*

$$\mathrm{CORR}[X,Y] \equiv \rho_{X,Y} = \frac{\mathrm{COV}[X,Y]}{\sqrt{\mathrm{VAR}[X]\mathrm{VAR}[Y]}} \tag{4.40}$$

On peut montrer que $-1 \leq \rho_{X,Y} \leq 1$. De plus, $\rho_{X,Y} = \pm 1$ si et seulement si l'on peut écrire que $Y = aX + b$, où $a \neq 0$. De façon plus précise, $\rho_{X,Y} = 1$ (respectivement -1) si $a > 0$ (respectivement $a < 0$). En fait, $\rho_{X,Y}$ est une mesure de *liaison linéaire* entre X et Y. Finalement, si X et Y sont des variables aléatoires indépendantes, alors on a: $\rho_{X,Y} = 0$.

Dans le cas où X et Y sont des variables aléatoires qui présentent une distribution normale, on a: $\rho_{X,Y} = 0 \Leftrightarrow X$ et Y sont indépendantes. Ce résultat est très important en pratique, car, si l'on *montre* (à l'aide d'un *test*) que deux variables aléatoires présentent (approximativement) une distribution normale et si l'on a trouvé que leur *coefficient de corrélation empirique* (voir le chapitre 7, page 399) est environ égal à zéro, alors, dans la cadre de procédures statistiques, on peut accepter qu'elles sont indépendantes.

Exemple 4.4.1. Considérons la fonction $p_{X,Y}$ donnée par le tableau suivant (voir l'exemple 4.1.1):

$y \backslash\, x$	-1	0	1	$p_Y(y)$
0	1/16	1/16	1/16	3/16
1	1/16	1/16	2/16	4/16
2	2/16	1/16	6/16	9/16
$p_X(x)$	4/16	3/16	9/16	1

À partir du tableau et des fonctions de probabilité marginales p_X et p_Y, on calcule

$$\begin{aligned}
E[X] &= -1 \times (4/16) + 0 \times (3/16) + 1 \times (9/16) = 5/16 \\
E[Y] &= 0 \times (3/16) + 1 \times (4/16) + 2 \times (9/16) = 22/16 \\
E[X^2] &= (-1)^2 \times (4/16) + 0^2 \times (3/16) + 1^2 \times (9/16) = 13/16 \\
E[Y^2] &= 0^2 \times (3/16) + 1^2 \times (4/16) + 2^2 \times (9/16) = 40/16 \\
E[XY] &= \sum_{x=-1}^{1} \sum_{y=0}^{2} xy\, p_{X,Y}(x,y) \\
&= 0 + (-1)(1)(1/16) + (-1)(2)(2/16) + 0 + 0 + 0 + 0 \\
&\quad +(1)(1)(2/16) + (1)(2)(6/16) \\
&= -1/16 - 4/16 + 2/16 + 12/16 = 9/16
\end{aligned}$$

Il s'ensuit que

$$\mathrm{VAR}[X] = E[X^2] - (E[X])^2 = \frac{13}{16} - \left(\frac{5}{16}\right)^2 = \frac{183}{(16)^2}$$

$$\mathrm{VAR}[Y] = E[Y^2] - (E[Y])^2 = \frac{40}{16} - \left(\frac{22}{16}\right)^2 = \frac{156}{(16)^2}$$

et

$$\mathrm{COV}[X,Y] = E[XY] - E[X]E[Y] = \frac{9}{16} - \left(\frac{5}{16}\right)\left(\frac{22}{16}\right) = \frac{34}{(16)^2}$$

Finalement, on calcule

$$\rho_{X,Y} \equiv \mathrm{CORR}[X,Y] = \frac{34/(16)^2}{\left(\frac{183}{(16)^2} \cdot \frac{156}{(16)^2}\right)^{1/2}} = \frac{34}{\sqrt{183 \cdot 156}} \simeq 0{,}2012$$

♢

Remarque. En général, il est plus avantageux de calculer les moyennes $E[X]$, $E[Y]$ et $E[XY]$ avant de calculer les moyennes des variables au carré. En effet, si $E[XY] - E[X]E[Y] = 0$, alors $\mathrm{CORR}[X, Y] = 0$ (si X et Y ne sont pas des constantes, de sorte que les variances de X et Y sont strictement positives).

Exemple 4.4.2. La fonction de densité conjointe du vecteur aléatoire continu (X, Y) est

$$f_{X,Y}(x, y) = \begin{cases} e^{-2y} \text{ si } 0 < x < 2,\ y > 0 \\ 0 \quad \text{ailleurs} \end{cases}$$

Quel est le coefficient de corrélation de X et Y?

Solution. On peut montrer que, lorsque la fonction de densité conjointe de (X, Y) peut être décomposée en un produit d'une fonction de x seulement et d'une fonction de y seulement, les variables aléatoires X et Y sont indépendantes, pourvu qu'il n'y ait pas de relation entre x et y dans l'ensemble $D_{X,Y}$ des valeurs possibles du couple (X, Y). C'est-à-dire que cet ensemble $D_{X,Y}$ est de la forme

$$D_{X,Y} = \{(x, y) \in \mathbb{R}^2 \colon c_1 < x < c_2, k_1 < y < k_2\}$$

où les c_i et les k_i sont des constantes pour $i = 1, 2$.

Ici, les valeurs possibles de X et de Y ne sont pas reliées, et l'on peut écrire que

$$e^{-2y} = g(x)h(y)$$

où $g(x) \equiv 1$ et $h(y) = e^{-2y}$. Donc, on peut conclure que X et Y sont indépendantes. Il s'ensuit que $\mathrm{CORR}[X, Y] = 0$.

On peut vérifier que X et Y sont effectivement indépendantes; on a:

$$f_X(x) = \int_0^\infty e^{-2y}\, dy = -\frac{1}{2} e^{-2y} \Big|_0^\infty = \frac{1}{2} \quad \text{si } 0 < x < 2$$

et

$$f_Y(y) = \int_0^2 e^{-2y}\, dx = 2e^{-2y} \quad \text{si } y > 0$$

Ainsi, on a bien: $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ pour tout couple (x, y). Notons que $X \sim \mathrm{U}(0, 2)$ et $Y \sim \mathrm{Exp}(\lambda = 2)$ dans cet exemple. ♢

4.5 Théorèmes limites

Dans cette section, nous allons donner deux théorèmes limites. Le premier est utile, en particulier, en statistique, tandis que le second est en fait le théorème le plus important des probabilités.

Théorème 4.5.1. (Loi des grands nombres) *Supposons que $X_1, X_2, \ldots$ sont des variables aléatoires indépendantes qui possèdent la même fonction de répartition que la variable X, dont la moyenne μ_X existe. Alors, pour n'importe quelle constante $\epsilon > 0$, on a:*

$$\lim_{n\to\infty} P\left[\left|\frac{X_1 + \ldots + X_n}{n} - \mu_X\right| > \epsilon\right] = 0 \tag{4.41}$$

Remarques.
(i) Ce théorème porte le nom, de façon plus précise, de loi *faible* des grands nombres. Il y a également la loi *forte* des grands nombres, pour laquelle l'espérance mathématique de $|X|$ doit exister.

(ii) En pratique, la moyenne μ_X de la variable aléatoire X est inconnue. Pour l'*estimer*, on prend plusieurs *observations* (indépendantes) X_i de X. Le résultat ci-dessus permet d'affirmer que la moyenne arithmétique de ces observations converge (*en probabilité*) vers la moyenne inconnue de X.

(iii) On écrit que les variables aléatoires $X_1, X_2, \ldots$ sont *i.i.d.* (indépendantes et identiquement distribuées).

Théorème 4.5.2. (Théorème central limite) *Supposons que $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes qui possèdent la même fonction de répartition que la variable X, dont la moyenne μ_X et la variance σ_X^2 existent ($\sigma_X > 0$). Alors la distribution de $S_n := \sum_{i=1}^n X_i$ tend vers celle d'une distribution normale, de moyenne $n\mu_X$ et variance $n\sigma_X^2$, lorsque n tend vers l'infini.*

Remarques.
(i) Soit

$$\bar{X} := \sum_{i=1}^{n} \frac{X_i}{n} \tag{4.42}$$

Alors on peut affirmer que la distribution de $\bar{X}$ tend vers celle d'une distribution $\mathrm{N}(\mu_X, \sigma_X^2/n)$.

(ii) En général, si l'on additionne 30 variables aléatoires indépendantes X_i ou plus, alors la distribution normale devrait être une bonne approximation de la distribution exacte (souvent inconnue) de cette somme. Toutefois, le nombre de variables qu'il faut additionner, pour obtenir une bonne approximation, dépend en fait du degré d'*asymétrie* de la distribution de X.

(iii) On peut, sous certaines conditions, généraliser le théorème central limite (TCL) au cas où les variables aléatoires $X_1, \ldots, X_n$ ne sont pas nécessairement identiquement distribuées. En effet, si la moyenne μ_{X_i} et la variance $\sigma_{X_i}^2$ de X_i existent pour tout i, alors, lorsque n est assez grand, on a:

$$\sum_{i=1}^{n} X_i \approx \mathrm{N}\left(\sum_{i=1}^{n} \mu_{X_i}, \sum_{i=1}^{n} \sigma_{X_i}^2\right) \tag{4.43}$$

et

$$\bar{X} \approx \mathrm{N}\left(\frac{1}{n}\sum_{i=1}^{n} \mu_{X_i}, \frac{1}{n^2}\sum_{i=1}^{n} \sigma_{X_i}^2\right) \tag{4.44}$$

Exemple 4.5.1. Une ville américaine compte 10.000 unités d'habitation et 2 usines. La demande d'eau potable (en gallons) de la part d'une unité d'habitation pour une journée quelconque est une variable aléatoire D telle que $E[D] = 50$ et $\mathrm{VAR}[D] = 400$. Dans le cas des usines, la demande d'eau potable présente (approximativement) une distribution $\mathrm{N}(10.000, (2000)^2)$ pour l'usine 1, et une distribution $\mathrm{N}(25.000, (5000)^2)$ pour l'usine 2. Soit D_i, pour $i = 1, \ldots, 10.000$, la demande d'eau potable de la part de la i^{e} unité d'habitation et U_i, pour $i = 1, 2$, la demande de la part de l'usine i. On suppose que les variables aléatoires D_i et U_i sont indépendantes et on pose:

$$X_d = \sum_{i=1}^{10.000} D_i \quad \text{(la demande domestique)}$$

et

$$X_t = X_d + U_1 + U_2 \quad \text{(la demande totale)}$$

(a) Trouver le nombre a tel que $P[X_d \geq a] \simeq 0{,}01$.

(b) Quelle devrait être la capacité de production de l'usine de filtration, si l'on veut satisfaire la demande totale avec une probabilité de 0,98?

Solution. (a) Par le théorème central limite, on peut écrire que

$$X_d \approx \mathrm{N}(10.000(50), 10.000(20^2))$$

Remarque. On suppose que les variables aléatoires D_i sont indépendantes entre elles.

Alors on a:

$$P[X_d \geq a] = 1 - P[X_d < a] \simeq 1 - P\left[Z < \frac{a - 500.000}{\sqrt{10.000}(20)}\right]$$

où $Z \sim \mathrm{N}(0,1)$. Il s'ensuit que

$$P[X_d \geq a] \simeq 0,01 \quad \Longleftrightarrow \quad P\left[Z < \frac{a - 500.000}{2000}\right] \simeq 0{,}99$$

Or, on trouve dans le tableau A.3, à la page 516, que $P[Z \leq 2{,}33] \simeq 0{,}99$. Donc, on a:

$$a \simeq 500.000 + 2000(2{,}33) = 504.660$$

(b) Par indépendance, on peut écrire que $X_t \approx \mathrm{N}(\mu, \sigma^2)$, où (voir la formule (4.29))

$$\mu = 500.000 + 10.000 + 25.000 = 535.000$$

et

$$\sigma^2 = (2000)^2 + (2000)^2 + (5000)^2 = 33.000.000$$

Soit c la capacité de l'usine de filtration. On cherche la valeur de c telle que $P[X_t \leq c] = 0{,}98$. Puisque $P[Z \leq 2{,}055] \overset{\text{tab. A.3}}{\simeq} 0{,}98$, en procédant comme en (a) on trouve que

$$c \simeq 535.000 + \sqrt{33.000.000}(2{,}055) \simeq 546.805$$

Remarque. On voit avec cet exemple qu'il n'est pas nécessaire de connaître la forme exacte de la fonction p_X ou f_X pour pouvoir appliquer le théorème central limite; il suffit de connaître la moyenne et la variance de X. ♢

4.6 Exercices du chapitre 4

Exercices résolus

Question n^{o} 1

Soit

$$p_{X,Y}(x,y) = \frac{1}{6} \quad \text{si } x = 0 \text{ ou } 1, \text{ et } y = 0, 1 \text{ ou } 2$$

Calculer $p_X(x)$.

Solution. On a:

$$p_X(x) = \sum_{y=0}^{2} p_{X,Y}(x,y) = \sum_{y=0}^{2} \frac{1}{6} = \frac{1}{2} \quad \text{si } x = 0 \text{ ou } 1$$

Question n^{o} 2

Calculer $f_X(x \mid Y = y)$ si

$$f_{X,Y}(x,y) = x + y \quad \text{pour } 0 < x < 1, 0 < y < 1$$

Solution. On trouve que

$$f_Y(y) = \int_0^1 (x+y)\,dx = \frac{1}{2} + y \quad \text{si } 0 < y < 1$$

Alors on peut écrire que

$$f_X(x \mid Y = y) = \frac{x+y}{\frac{1}{2}+y} \quad \text{si } 0 < x < 1 \text{ et } 0 < y < 1$$

Question n^{o} 3

Supposons que X et Y sont deux variables aléatoires telles que $E[X] = E[Y] = 0$, $E\left[X^2\right] = E\left[Y^2\right] = 1$ et $\rho_{XY} = 1$. Calculer $\mathrm{COV}[X,Y]$.

Solution. On trouve d'abord que $\mathrm{VAR}[X] = \mathrm{VAR}[Y] = 1 - 0^2 = 1$. Alors

$$\mathrm{COV}[X,Y] = \rho_{X,Y}\sigma_X\sigma_Y = 1 \cdot 1 \cdot 1 = 1$$

Question n^{o} 4

Calculer $P\left[X+Y>1\right]$ si

$$f_{X,Y}\left(x,y\right)=1 \quad \text{pour } 0<x<1, 0<y<1$$

Solution. On peut écrire que $P[X+Y>1]=1/2$, par symétrie. On peut vérifier ce résultat comme suit:

$$P[X+Y>1]=\int_0^1\int_{1-x}^1 1\,dydx=\int_0^1\left[1-(1-x)\right]dx=\frac{1}{2}$$

Question n^{o} 5

Supposons que

$$p_{X,Y}\left(x,y\right)=\frac{8}{9}\left(\frac{1}{2}\right)^{x+y} \quad \text{si } x=0 \text{ ou } 1, \text{ et } y=1 \text{ ou } 2$$

Calculer $E\left[XY\right]$.

Solution. On a:

$$E[XY]=(1)(1)\left(\frac{8}{9}\right)\left(\frac{1}{2}\right)^{1+1}+(1)(2)\left(\frac{8}{9}\right)\left(\frac{1}{2}\right)^{1+2}=\left(\frac{8}{9}\right)\left(\frac{1}{4}+\frac{1}{4}\right)=\frac{4}{9}$$

Question n^{o} 6

Supposons que X et Y sont deux variables aléatoires telles que $\mathrm{VAR}[X]=\mathrm{VAR}[Y]=1$ et $\mathrm{COV}[X,Y]=1$. Calculer $\mathrm{VAR}[X-2Y]$.

Solution. On a:

$$\mathrm{VAR}[X-2Y]=\mathrm{VAR}[X]+4\,\mathrm{VAR}[Y]-4\,\mathrm{COV}[X,Y]=1+4-4(1)=1$$

Question n^{o} 7

Soit $X\sim \mathrm{N}(0,1)$, $Y\sim \mathrm{N}(1,2)$ et $Z\sim \mathrm{N}(3,4)$ des variables aléatoires indépendantes. Quelle distribution présente $W:=X-Y+2Z$? Donner aussi le ou les paramètres de cette distribution.

Solution. On peut écrire que $W\sim \mathrm{N}(0-1+2(3),1+2+4(4))\equiv \mathrm{N}(5,19)$.

Question n^{o} 8

Supposons que $X\sim \mathrm{Poi}(\lambda=100)$. Quelle autre distribution de probabilité peut-on utiliser pour calculer approximativement $p:=P\left[X\leq 100\right]$? Donner

aussi le ou les paramètres de cette distribution ainsi que la valeur approximative de p.

Solution. Par le théorème central limite, on peut écrire que

$$P[\text{Poi}(100) \leq 100] \simeq P\left[\text{N}(100,100) \leq 100 + \frac{1}{2}\right] = \Phi(0{,}05) \overset{\text{tab. A.3}}{\simeq} 0{,}5199$$

car si $X \sim \text{Poi}(100)$, alors X possède la même distribution que $\sum_{i=1}^{100} X_i$, où les X_i sont des variables aléatoires indépendantes qui présentent une distribution de Poisson de paramètre 1. Donc, on peut se servir de la distribution $\text{N}(100,100)$.

Remarque. Notons que ce ne sont pas tous les auteurs qui font une correction de continuité dans ce cas (comme dans le cas de la distribution binomiale). Si l'on ne fait pas de correction de continuité, alors on obtient directement que

$$P[\text{Poi}(100) \leq 100] \simeq P\left[\text{N}(100,100) \leq 100\right] = \Phi(0) = \frac{1}{2}$$

En fait, on trouve, avec un logiciel, que $P[\text{Poi}(100) \leq 100] \simeq 0{,}5266$. Donc, ici le fait de faire une correction de continuité améliore l'approximation.

Question n$^{\text{o}}$ 9

Supposons que X présente une distribution $\text{B}(n = 100, p = 0{,}4)$. Utiliser la distribution $\text{N}(40,24)$ pour calculer approximativement $P\left[X = 40\right]$.

Solution. On a:

$$\begin{aligned} P[X = 40] &= P[39{,}5 \leq X \leq 40{,}5] \simeq P[39{,}5 \leq \text{N}(40,24) \leq 40{,}5] \\ &= 2\,\Phi\left(\frac{40{,}5 - 40}{\sqrt{24}}\right) - 1 \simeq 2\,\Phi(0{,}10) - 1 \overset{\text{tab. A.3}}{\simeq} 0{,}0796 \end{aligned}$$

Ou encore:

$$\begin{aligned} P[X = 40] &\simeq f_Y(40) \quad \text{où } Y \sim \text{N}(40,24) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{24}} \exp\left\{-\frac{1}{2}\frac{(40-40)^2}{24}\right\} = \frac{1}{\sqrt{2\pi}\sqrt{24}} \simeq 0{,}0814 \end{aligned}$$

Remarque. La réponse obtenue en utilisant la distribution binomiale et un logiciel est $P[X = 40] \simeq 0{,}0812$, ce qui est également la valeur que l'on trouve en calculant (avec plus de précision que ci-dessus) $\Phi((40{,}5-40)/\sqrt{24}) \simeq \Phi(0{,}102)$ plutôt que $\Phi(0{,}10)$.

Question n° 10

On définit $Y = \sum_{i=1}^{50} X_i$, où $E[X_i] = 0$ pour $i = 1, \ldots, 50$ et les X_i sont des variables aléatoires continues et indépendantes. Calculer approximativement $P[Y \geq 0]$.

Solution. Par le théorème central limite, on peut écrire que

$$Y := \sum_{i=1}^{50} X_i \approx \mathrm{N}(50(0), \sigma_Y^2)$$

de sorte que $P[Y \geq 0] \simeq P[\mathrm{N}(0, \sigma_Y^2) \geq 0] = 1/2$.

Question n° 11

La fonction de probabilité conjointe, $p_{X,Y}$, du couple (X, Y) est donnée par le tableau suivant:

$y \backslash x$	-1	0	1
0	1/9	1/9	1/9
2	2/9	2/9	2/9

(a) Les variables aléatoires X et Y sont-elles indépendantes? Justifier.

(b) Calculer $F_{X,Y}\left(0, \frac{1}{2}\right)$.

(c) Soit $Z = X^4$; calculer $p_Z(z)$.

(d) Calculer $E\left[X^2 Y^2\right]$.

Solution. (a) On calcule d'abord $p_X(x) \equiv 1/3$, $p_Y(0) = 1/3$ et $p_Y(2) = 2/3$. On peut vérifier que $p_X(x) p_Y(y) = p_{X,Y}(x, y)\ \forall (x, y)$. Donc, X et Y sont indépendantes.

(b) $F_{X,Y}\left(0, \frac{1}{2}\right) = P\left[X \leq 0, Y \leq \frac{1}{2}\right] = p_{X,Y}(-1, 0) + p_{X,Y}(0, 0) = 2/9$.

(c) D'après (a), on a: $p_X(x) = 1/3$ pour $x = -1, 0, 1$; alors

z	0	1
$p_Z(z)$	1/3	2/3

(d) On a: $E[X^2 Y^2] = (-1)^2 (2)^2 \frac{2}{9} + (1)^2 (2)^2 \frac{2}{9} + 0 = \frac{16}{9}$.

Question n° 12

Soit

$$f_{X,Y}(x, y) = \begin{cases} 2 \text{ si } 0 < x < y, 0 < y < 1 \\ 0 \text{ ailleurs} \end{cases}$$

Calculer $P\left[X \geq Y^2\right]$.

Solution. On a (voir la figure 4.3):

$$P[X \geq Y^2] = \int_0^1 \int_x^{\sqrt{x}} 2\, dydx = \int_0^1 2(\sqrt{x} - x)dx$$
$$= 2\left(\frac{x^{3/2}}{3/2} - \frac{x^2}{2}\right)\Bigg|_0^1 = \frac{1}{3}$$

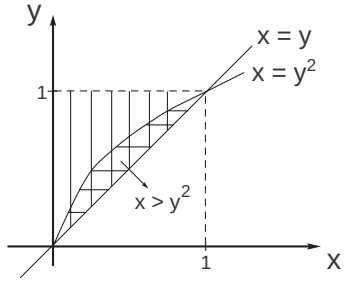

Fig. 4.3. Figure pour l'exercice résolu n° 12

Question n° 13

Les autobus de la ville passent à un certain coin de rue, entre 7 h et 19 h 30, selon un processus de Poisson à la cadence moyenne de quatre par heure. Soit $Y = \sum_{k=1}^{50} X_k$, où X_k est le nombre d'autobus qui passent pendant la k^{e} période de 15 minutes à partir de 7 h.

(a) Quelle est la distribution *exacte* de Y et son ou ses paramètres?

(b) Quelle autre distribution de probabilité peut approcher la distribution de Y? Justifier votre réponse et donner également les paramètres de cette distribution.

Solution. (a) Soit $N(t)$ le nombre d'autobus dans l'intervalle $[0, t]$, où t est en heures. On peut écrire que $Y \equiv N(t = 12{,}5) \sim \text{Poi}(50)$.

(b) On a: $X_k \sim \text{Poi}(1)$, de sorte que $E[X_k] = \text{VAR}[X_k] = 1\ \forall k$. Alors, étant donné que les X_k sont des variables aléatoires indépendantes, on déduit du théorème central limite que $Y \approx \text{N}(50, 50)$.

Question n° 14

On considère la variable aléatoire discrète X dont la fonction de probabilité est donnée par

x	0	1	2
$p_X(x)$	1/2	1/4	1/4

Supposons que X_1 et X_2 sont deux variables aléatoires indépendantes qui possèdent la même distribution que X. Calculer $P[X_1 = X_2]$.

Solution. On a:

$$P[X_1 = X_2] = \sum_{i=0}^{2} P[\{X_1 = i\} \cap \{X_2 = i\}] \overset{\text{ind.}}{=} \sum_{i=0}^{2} P[X_1 = i]P[X_2 = i]$$
$$= (1/2)^2 + (1/4)^2 + (1/4)^2 = 3/8 = 0{,}375$$

Question n^o 15

Le tableau ci-dessous donne la fonction $p_{X,Y}(x, y)$ du couple de variables aléatoires discrètes (X, Y):

$y\backslash x$	0	1	3	4
1	0,1	0,1	0	0,2
2	0,3	0	0,2	0,1

Calculer $P[\{X < 5\} \cap \{Y < 2\}]$.

Solution. On a:

$$P[\{X < 5\} \cap \{Y < 2\}] = P[\{X \leq 4\} \cap \{Y \leq 1\}] = P[Y = 1]$$
$$= 0{,}1 + 0{,}1 + 0{,}2 = 0{,}4$$

Question n^o 16

Calculer la covariance de X_1 et X_2 si

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} 2 - x_1 - x_2 \text{ pour } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

Solution. On calcule d'abord

$$f_{X_1}(x_1) = \int_0^1 (2 - x_1 - x_2)\, dx_2 = \frac{3}{2} - x_1 \quad \text{si } 0 < x_1 < 1$$

Par symétrie, on peut alors écrire que

$$f_{X_2}(x_2) = \frac{3}{2} - x_2 \quad \text{si } 0 < x_2 < 1$$

On calcule ensuite

$$E[X_i] = \int_0^1 x_i \left(\frac{3}{2} - x_i\right) dx_i = \frac{5}{12} \quad \text{pour } i = 1, 2$$

et

$$E[X_1 X_2] = \int_0^1 \int_0^1 x_1 x_2 \left(2 - x_1 - x_2\right) dx_1 dx_2 = \frac{1}{6}$$

Il s'ensuit que $\mathrm{COV}[X_1, X_2] = E[X_1 X_2] - E[X_1]E[X_2] = -\frac{1}{144} \simeq -0{,}0069$.

Question n$^{\text{o}}$ 17

Supposons que $Y = 1/X$, où X est une variable aléatoire discrète telle que

x	1	2
$p_X(x)$	1/3	2/3

On définit $W = Y_1 - Y_2$, où Y_1 et Y_2 sont deux variables aléatoires indépendantes et distribuées de façon identique à Y. Calculer $p_W(w)$.

Solution. On trouve d'abord que

y	1/2	1
$p_Y(y)$	2/3	1/3

ce qui implique que

w	$-1/2$	0	1/2
$p_W(w)$	2/9	5/9	2/9

car $P[W = -1/2] = P[Y_1 = 1/2, Y_2 = 1] \overset{\text{ind.}}{=} (2/3)(1/3) = 2/9$, etc.

Question n$^{\text{o}}$ 18

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes, où X_i présente une distribution exponentielle de paramètre $\lambda = 2$ pour $i = 1, \ldots, n$. Utiliser le théorème central limite pour trouver la valeur de n pour laquelle

$$P\left[\sum_{i=1}^{n} X_i > \frac{n}{2} + 1\right] \simeq 0{,}4602$$

Solution. On peut écrire que $\sum_{i=1}^{n} X_i \approx \mathrm{N}\big(n\left(\frac{1}{2}\right), n\left(\frac{1}{4}\right)\big)$, car $\mu_{X_i} = 1/2$ et $\sigma^2_{X_i} = 1/4$ pour tout i. Alors

$$P\left[\sum_{i=1}^{n} X_i > \frac{n}{2}+1\right] \simeq P\left[\mathrm{N}(0,1) > \frac{\frac{n}{2}+1-\frac{n}{2}}{\sqrt{n}/2}\right] \simeq 0{,}4602$$

$$\iff \quad \Phi\left(\frac{2}{\sqrt{n}}\right) \simeq 0{,}5398 \quad \overset{\text{tab. A.3}}{\iff} \quad \frac{2}{\sqrt{n}} \simeq 0{,}1 \quad \iff \quad n \simeq 400$$

Question n$^{\text{o}}$ 19

Un autobus passe à un certain coin de rue tous les matins vers 9 h. Soit X la différence (en minutes) entre l'instant où l'autobus passe et 9 h. On suppose que X présente approximativement une distribution $\mathrm{N}\big(\mu = 0, \sigma^2 = 25\big)$. On considère deux journées indépendantes. Soit X_k la valeur de la variable aléatoire X lors de la k^{e} journée, pour $k = 1, 2$.

(a) Calculer la probabilité $P\,[X_1 - X_2 > 15]$.

(b) Donner la fonction de densité conjointe $f_{X_1,X_2}\,(x_1, x_2)$.

(c) Calculer (i) $P\,[X_1 = 2 \mid X_1 > 1]$ et (ii) $P\,[X_1 < 2 \mid X_1 = 1]$.

Solution. (a) On a: $X_1 - X_2 \sim \mathrm{N}(0 - 0, 25 + 25) \equiv \mathrm{N}(0, 50)$. On calcule alors

$$P\,[X_1 - X_2 > 15] \quad = \quad P\left[\mathrm{N}(0,1) > \frac{15-0}{\sqrt{50}}\right] \simeq 1 - \Phi(2{,}12)$$
$$\overset{\text{tab. A.3}}{\simeq} 1 - 0{,}9830 = 0{,}0170$$

(b) Par indépendance, on peut écrire que

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \frac{1}{(2\pi)(25)} \exp\left\{-\frac{(x_1^2 + x_2^2)}{50}\right\}$$

pour $(x_1, x_2) \in \mathbb{R}^2$.

(c) (i) On a: $P\,[X_1 = 2 \mid X_1 > 1] = 0$, car X_1 est une variable aléatoire *continue*.

(ii) On peut écrire que $P\,[X_1 < 2 \mid X_1 = 1] = 1$, car $\{X_1 = 1\} \subset \{X_1 < 2\}$.

Question n$^{\text{o}}$ 20

Un assemblage comprend 100 sections. La longueur de chaque section (en centimètres) est une variable aléatoire dont la moyenne égale 10 et la variance égale 0,9. De plus, les sections sont indépendantes. La norme pour la longueur totale de l'assemblage est de 1000 cm ± 30 cm. Quelle est approximativement la probabilité que l'assemblage ne respecte pas la norme en question?

Solution. Soit X_i la longueur de la i^{e} section, pour $i = 1, \ldots, 100$. Par le théorème central limite, on peut écrire que

$$P\left[970 \leq \sum_{i=1}^{100} X_i \leq 1030\right] \simeq P\left[970 \leq \mathrm{N}(\mu, \sigma^2) \leq 1030\right]$$

où $\mu = 100 \times 10 = 1000$, et $\sigma^2 = 100 \times 0,9 = 90$. Donc, on cherche environ

$$1 - [\Phi(3{,}16) - \Phi(-3{,}16)] \overset{\text{tab. A.3}}{\simeq} 1 - 2 \times (0{,}9992 - 1) = 0{,}0016$$

Question n° 21

Soit

$$f_{X,Y}(x, y) = \begin{cases} 3x^2 e^{-x} y\,(1-y) \text{ si } x > 0, 0 < y < 1 \\ 0 \qquad\qquad \text{ailleurs} \end{cases}$$

(a) Calculer $f_X(x)$ et $f_Y(y)$. Quelle est la distribution de X et celle de Y?

(b) X et Y sont-elles des variables aléatoires indépendantes? Justifier.

(c) Calculer le coefficient d'aplatissement de X.

(d) Calculer le coefficient d'asymétrie de Y.

Solution. (a) On a:

$$\begin{aligned} f_X(x) &= \int_0^1 3x^2 e^{-x} y(1-y)\,dy = 3x^2 e^{-x} \int_0^1 y(1-y)\,dy \\ &= 3x^2 e^{-x} \left(\frac{y^2}{2} - \frac{y^3}{3}\right)\Big|_0^1 = \frac{x^2 e^{-x}}{2} \quad \text{si } x > 0 \end{aligned}$$

On trouve que $X \sim \mathrm{G}(\alpha = 3, \lambda = 1)$.

Ensuite, on calcule

$$\begin{aligned} f_Y(y) &= \int_0^\infty 3x^2 e^{-x} y(1-y)\,dx = 3y(1-y) \int_0^\infty x^2 e^{-x}\,dx \\ &= 3y(1-y)\Gamma(3) = 6y(1-y) \quad \text{si } 0 < y < 1 \end{aligned}$$

Dans ce cas, on trouve que $Y \sim \mathrm{Be}(\alpha = 2, \beta = 2)$.

(b) On peut vérifier que $f_{X,Y}(x, y) \equiv f_X(x) f_Y(y)$. Donc, par définition, X et Y sont des variables aléatoires indépendantes.

(c) On a: $\mathrm{VAR}[X] \overset{\text{(a)}}{=} 3/(1)^2 = 3$. On calcule ensuite

$$E[X^k] = \int_0^\infty \frac{1}{2} x^{2+k} e^{-x}\,dx = \frac{1}{2}\Gamma(k+3)$$

pour $k = 1, 2, \ldots$ Alors on a:

$$\begin{aligned}\beta_2 &= \frac{E[X^4] - 4E[X^3]E[X] + 6E[X^2](E[X])^2 - 4E[X](E[X])^3 + (E[X])^4}{(\mathrm{VAR}[X])^2} \\ &= \frac{\frac{6!}{2} - 4\left(\frac{5!}{2}\right)\left(\frac{3!}{2}\right) + 6\left(\frac{4!}{2}\right)\left(\frac{3!}{2}\right)^2 - 3\left(\frac{3!}{2}\right)^4}{(3)^2} \\ &= \frac{360 - 720 + 648 - 324 + 81}{9} = \frac{45}{9} = 5\end{aligned}$$

(d) On calcule d'abord

$$E[Y] = \int_0^1 y\, 6y(1-y)\, dy = \int_0^1 6y^2(1-y)\, dy = 6\left(\frac{1}{3} - \frac{1}{4}\right) = \frac{1}{2}$$

Notons que

$$f_Y\left(\tfrac{1}{2} - y\right) = f_Y\left(\tfrac{1}{2} + y\right)$$

pour tout $y \in (0, \frac{1}{2})$. C'est-à-dire que la fonction $f_Y(y)$ est symétrique par rapport à la moyenne de Y. Puisque toutes les valeurs prises par la variable aléatoire Y sont situées dans un intervalle borné, on peut affirmer que $\beta_1 = 0$.

Question n$^{\mathrm{o}}$ 22

Le tableau suivant donne la fonction de probabilité conjointe $p_{X,Y}(x, y)$ du couple (X, Y):

$y \backslash x$	0	1	2
-1	1/9	0	1/9
0	2/9	0	2/9
1	0	1/3	0

(a) Calculer $p_X(x)$ et $p_Y(y)$.

(b) X et Y sont-elles des variables aléatoires indépendantes? Justifier.

(c) Calculer (i) $p_Y(y \mid X = 1)$ et (ii) $p_Y(y \mid X \leq 1)$.

(d) Calculer le coefficient de corrélation de X et Y.

(e) Soit $W = \max\{X, Y\}$. Calculer $p_W(w)$.

Solution. (a) En additionnant les éléments des colonnes et des lignes du tableau, respectivement, on trouve que

x	0	1	2
$p_X(x)$	1/3	1/3	1/3

et

y	-1	0	1
$p_Y(y)$	2/9	4/9	1/3

(b) On a, par exemple: $p_X(0)p_Y(-1) = \frac{1}{3} \times \frac{2}{9} \neq \frac{1}{9} = p_{X,Y}(0,-1)$. Donc, X et Y ne sont pas des variables aléatoires indépendantes.

Remarque. Étant donné qu'il y a des "0" dans le tableau, X et Y ne pouvaient pas être indépendantes. En effet, un "0" dans le tableau ne sera jamais égal au produit de la somme des éléments de la ligne et de la colonne correspondantes.

(c) (i) Par définition, on a:

$$p_Y(y \mid X = 1) = \frac{p_{X,Y}(1,y)}{p_X(1)} \overset{\text{(a)}}{=} 3\, p_{X,Y}(1,y) = \begin{cases} 0 \text{ si } y = -1 \\ 0 \text{ si } y = 0 \\ 1 \text{ si } y = 1 \end{cases}$$

C'est-à-dire que $Y \mid \{X = 1\}$ est la constante 1.

(ii) On trouve d'abord que $P[X \leq 1] \overset{\text{(a)}}{=} \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$. Il s'ensuit que

$$p_Y\left(-1 \mid X \leq 1\right) = \frac{3}{2} P[\{Y = -1\} \cap \{X \leq 1\}] = \frac{3}{2}\left(\frac{1}{9} + 0\right) = \frac{1}{6}$$

De même, on calcule $p_Y\left(0 \mid X \leq 1\right) = \frac{3}{2}\left(\frac{2}{9} + 0\right) = \frac{1}{3}$. Donc, on a:

y	-1	0	1	Σ
$p_Y(y \mid X \leq 1)$	1/6	1/3	1/2	1

(d) On calcule d'abord

$$E[X] \overset{\text{(a)}}{=} 1 \times \frac{1}{3} + 2 \times \frac{1}{3} = 1, \quad E[Y] \overset{\text{(a)}}{=} -1 \times \frac{2}{9} + 1 \times \frac{1}{3} = \frac{1}{9}$$

et

$$E[XY] = 1 \times 1 \times \frac{1}{3} + 2 \times (-1) \times \frac{1}{9} = \frac{1}{9}$$

Il s'ensuit que

$$\mathrm{COV}[X,Y] = E[XY] - E[X]E[Y] = \frac{1}{9} - 1 \times \frac{1}{9} = 0$$

et, par conséquent, $\mathrm{CORR}[X,Y] = 0$ (car $\mathrm{STD}[X] > 0$ et $\mathrm{STD}[Y] > 0$).

(e) On a:

$$W = \begin{cases} 0 \text{ si } (X,Y) = (0,-1) \text{ ou } (0,0) \\ 1 \text{ si } (X,Y) = (0,1) \text{ ou } (1,-1) \text{ ou } (1,0) \text{ ou } (1,1) \\ 2 \text{ si } (X,Y) = (2,-1) \text{ ou } (2,0) \text{ ou } (2,1) \end{cases}$$

En utilisant la fonction $p_{X,Y}(x,y)$, on trouve, par incompatibilité, que la fonction $p_W(w)$ est donnée par

w	0	1	2	Σ
$p_W(w)$	1/3	1/3	1/3	1

Question n° 23

On considère un couple de variables aléatoires discrètes (X,Y) dont la fonction de probabilité conjointe $p_{X,Y}(x,y)$ est donnée par

$y\backslash x$	1	2	3
2	1/12	1/6	1/12
3	1/6	0	1/6
4	0	1/3	0

Calculer $P[X+Y \leq 4 \mid X \leq 2]$.

Solution. On a:

$$P[X+Y \leq 4 \mid X \leq 2] = \frac{P[\{X+Y \leq 4\} \cap \{X \leq 2\}]}{P[X \leq 2]}$$
$$= \frac{P[(X,Y) \in \{(1,2),(1,3),(2,2)\}]}{1-\left(\frac{1}{12}+\frac{1}{6}+0\right)} = \frac{\frac{1}{12}+\frac{1}{6}+\frac{1}{6}}{\frac{3}{4}} = \frac{5}{9} = 0{,}\bar{5}$$

Question n° 24

Utiliser une distribution normale pour calculer approximativement la probabilité que, parmi 10.000 chiffres aléatoires (indépendants), le chiffre "7" apparaisse plus de 968 fois.

Solution. Soit X le nombre de fois que le chiffre "7" apparaît parmi les 10.000 chiffres aléatoires. Alors $X \sim \mathrm{B}(n=10.000, p=0{,}1)$. On cherche

$$P[X > 968] = P[X \geq 969] \simeq P\left[\mathrm{N}(0,1) \geq \frac{969-0{,}5-1000}{\sqrt{900}}\right]$$
$$= Q(-1{,}05) = \Phi(1{,}05) \overset{\text{tab. A.3}}{\simeq} 0{,}8531$$

Question n° 25

On prend un nombre X au hasard sur l'intervalle $(0,1)$ et un nombre Y au hasard sur l'intervalle $(0,X]$, de sorte que

$$f_{X,Y}(x,y) = \begin{cases} 1/x \text{ si } 0 < x < 1, 0 < y \leq x \\ 0 \quad \text{ailleurs} \end{cases}$$

(a) Montrer que

$$E\left[X^r Y^s\right] = \frac{1}{(s+1)(r+s+1)}$$

pour $r, s = 0, 1, 2, \ldots$

(b) Vérifier la formule en (a) lorsque $r = 2$ et $s = 0$ en calculant directement $E\left[X^2\right]$.

(c) Utiliser la partie (a) pour calculer le coefficient de corrélation de X et Y.

Solution. (a) On a:

$$\begin{aligned} E[X^r Y^s] &= \int_0^1 \int_0^x x^r y^s \frac{1}{x}\, dydx = \int_0^1 x^{r-1} \left.\frac{y^{s+1}}{s+1}\right|_0^x dx \\ &= \int_0^1 \frac{x^{r+s}}{s+1}\, dx = \frac{1}{(s+1)(r+s+1)} \quad \text{pour } r, s = 0, 1, 2, \ldots \end{aligned}$$

(b) On calcule

$$E[X^2] = \int_0^1 x^2 \cdot 1\, dx = \frac{1}{3}$$

et

$$E[X^2 Y^0] = \frac{1}{(0+1)(2+0+1)} = \frac{1}{3}$$

(c) On déduit de (a) que $E[X] = E[X^1 Y^0] = \frac{1}{2}$, $E[X^2] = E[X^2 Y^0] = \frac{1}{3}$, $E[Y] = \frac{1}{4}$, $E[Y^2] = \frac{1}{9}$ et $E[XY] = \frac{1}{6}$. Alors on a:

$$\text{VAR}[X] = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}, \quad \text{VAR}[Y] = \frac{1}{9} - \left(\frac{1}{4}\right)^2 = \frac{7}{144}$$

et

$$\rho_{X,Y} = \frac{\frac{1}{6} - \frac{1}{2} \cdot \frac{1}{4}}{\sqrt{\frac{1}{12} \cdot \frac{7}{144}}} = \sqrt{\frac{3}{7}} \simeq 0{,}6547$$

Question n° 26

Le tableau suivant donne une partie de la fonction $p_{X,Y}(x,y)$ du couple de variables aléatoires discrètes (X, Y):

$y \setminus x$	0	1	2	$p_Y(y)$
-1	1/16		1/16	1/4
0				1/2
1		0		1/4
$p_X(x)$	1/4			1

On a aussi:

y	-1	0	1
$p_Y(y \mid X = 2)$	1/8	3/8	1/2

(a) Trouver $P[X = 2]$.

(b) Compléter le tableau de la fonction $p_{X,Y}(x, y)$.

(c) On pose $W = Y + 1$. La distribution de W est alors un cas particulier d'une des distributions discrètes vues au chapitre 3. Trouver cette distribution et donner son ou ses paramètres.

Solution. (a) On a:

$$\frac{1}{8} = p_Y(-1 \mid X = 2) = \frac{p_{X,Y}(2,-1)}{p_X(2)} = \frac{1/16}{P[X=2]} \implies P[X = 2] = \frac{1}{2}$$

(b) On trouve d'abord que

$$p_{X,Y}(2,0) = p_Y(0 \mid X = 2)p_X(2) \overset{\text{(a)}}{=} (3/8)(1/2) = 3/16$$

De même, on a: $p_{X,Y}(2,1) = (1/2)(1/2) = 1/4$. On obtient alors le tableau suivant:

$y \setminus x$	0	1	2	$p_Y(y)$
-1	1/16	1/8	1/16	1/4
0	3/16	1/8	3/16	1/2
1	0	0	1/4	1/4
$p_X(x)$	1/4	1/4	1/2	1

(c) On peut vérifier que W présente une distribution binomiale de paramètres $n = 2$ et $p = 1/2$, car

$$p_W(w) = \binom{2}{w}(1/2)^2 \quad \text{pour } w = 0, 1, 2$$

Question n° 27

La fonction de densité conjointe du couple de variables aléatoires continues (X, Y) est donnée par

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}xy \text{ si } 0 < y < x < 2 \\ 0 \quad \text{ailleurs} \end{cases}$$

(a) Calculer $E[1/XY]$.

(b) Calculer $E\left[X^2\right]$.

(c) Quelle est la médiane, x_m, de la variable aléatoire X?

Solution. (a) On a:

$$E\left[\frac{1}{XY}\right] = \int_0^2 \int_0^x \frac{1}{xy}\frac{xy}{2}\,dydx = \int_0^2 \frac{x}{2}\,dx = 1$$

(b) On calcule d'abord (voir la figure 4.4)

$$f_X(x) = \int_0^x \frac{xy}{2}\,dy = \frac{x^3}{4} \quad \text{pour } 0 < x < 2$$

Alors

$$E[X^2] = \int_0^2 x^2\,\frac{x^3}{4}\,dx = \left.\frac{x^6}{24}\right|_0^2 = \frac{8}{3} \simeq 2{,}67$$

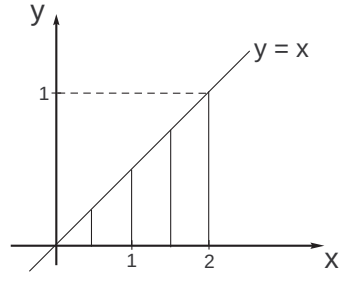

Fig. 4.4. Figure pour l'exercice résolu n° 27

(c) En utilisant la partie (b), on peut écrire que l'on cherche le nombre x_m pour lequel

$$\int_0^{x_m} \frac{x^3}{4}\,dx = \frac{1}{2} \iff \frac{x_m^4}{16} = \frac{1}{2} \implies x_m \simeq 1{,}68$$

car la médiane x_m doit être positive, puisque $X \in (0,2)$.

Question n$^{\text{o}}$ 28

Un appareil est constitué de deux composants indépendants placés en parallèle. La durée de vie X (en années) du composant n$^{\text{o}}$ 1 présente une distribution exponentielle de paramètre $\lambda = 1/2$, tandis que la durée de vie Y (en années) du composant n$^{\text{o}}$ 2 présente une distribution de Weibull de paramètres $\lambda = 2$ et $\beta = 2$; c'est-à-dire que

$$f_Y(y) = 4ye^{-2y^2} \quad \text{pour } y > 0$$

Calculer la probabilité que l'appareil dure moins d'un an.

Solution. On cherche

$$P[\{X < 1\} \cap \{Y < 1\}] \overset{\text{ind.}}{=} P[X < 1]P[Y < 1] \simeq 0{,}3402$$

car

$$P[X < 1] = \int_0^1 \frac{1}{2} e^{-x/2}\,dx = 1 - e^{-1/2}$$

et

$$P[Y < 1] = \int_0^1 4ye^{-2y^2}\,dy = 1 - e^{-2}$$

Question n$^{\text{o}}$ 29

On prend 100 nombres au hasard dans l'intervalle $[0,1]$. Soit S la somme des 100 nombres. Utiliser le théorème central limite pour calculer approximativement la probabilité $P\,[45 \leq S < 55]$.

Solution. Soit X_k le k^{e} nombre pris au hasard. On a: $E[X_k] = 1/2$ et $\text{VAR}[X_k] = 1/12$ pour tout k. Alors, par le théorème central limite, on peut écrire que

$$\begin{aligned} P\,[45 \leq S < 55] &\simeq P\left[45 \leq \text{N}\left(\tfrac{100}{2}, \tfrac{100}{12}\right) < 55\right] \\ &\simeq P[-1{,}73 \leq \text{N}(0,1) < 1{,}73] = \Phi(1{,}73) - \Phi(-1{,}73) \\ &= 2\,\Phi(1{,}73) - 1 \overset{\text{tab. A.3}}{\simeq} 2(0{,}958) - 1 = 0{,}916 \end{aligned}$$

Question n° 30

Le nombre d'inondations qui se produisent dans une certaine région au cours d'une année est une variable aléatoire qui présente une distribution de Poisson de paramètre $\alpha = 2$, indépendamment d'une année à l'autre. De plus, le temps (en jours) pendant lequel les terrains sont inondés, lors d'une inondation quelconque, est une variable aléatoire exponentielle de paramètre $\lambda = 1/5$. On suppose que les durées des inondations sont indépendantes. Utiliser le théorème central limite pour calculer (approximativement) la probabilité

(a) qu'au cours des 50 prochaines années, il se produise au moins 80 inondations dans cette région;

(b) que le temps total pendant lequel les terrains seront inondés au cours des 50 prochaines inondations soit inférieur à 200 jours.

Solution. (a) Soit X_i le nombre d'inondations au cours de la i^{e} année. Alors $X_i \sim \text{Poi}(\alpha = 2)\ \forall i$ et les X_i sont indépendantes. On cherche

$$P\left[\sum_{i=1}^{50} X_i \geq 80\right] \overset{\text{TCL}}{\simeq} P[\text{N}(50(2), 50(2)) \geq 80]$$
$$= P\left[\text{N}(0,1) \geq \frac{80-100}{10}\right] = 1 - \Phi(-2) = \Phi(2) \overset{\text{tab. A.3}}{\simeq} 0{,}9772$$

(b) Soit Y_i la durée de la i^{e} inondation. Alors $Y_i \sim \text{Exp}(\lambda = 1/5)\ \forall i$ et les Y_i sont des variables aléatoires indépendantes. Puisque $E[Y_i] = 1/\lambda = 5$ et $\text{VAR}[Y_i] = 1/\lambda^2 = 25$, on calcule

$$P\left[\sum_{i=1}^{50} Y_i < 200\right] \overset{\text{TCL}}{\simeq} P[\text{N}(50(5), 50(25)) < 200]$$
$$\simeq P\left[\text{N}(0,1) < -1{,}41\right] \overset{\text{tab. A.3}}{\simeq} 1 - 0{,}9207 \simeq 0{,}079$$

Exercices

Question n° 1

Des appels à un central téléphonique arrivent selon un processus de Poisson de taux λ par minute. On sait, par expérience, que la probabilité de recevoir exactement un appel durant une minute est égale au triple de la probabilité

de ne recevoir aucun appel durant la même période. On considère 100 périodes consécutives d'une minute et on désigne par U le nombre de périodes où aucun appel n'a été reçu.

(a) Utiliser une approximation par une distribution normale pour calculer $P[U = 5]$.

(b) Utiliser le théorème central limite pour calculer approximativement

$$P\left[\frac{1}{100}\sum_{i=1}^{100} X_i \geq 3{,}1\right]$$

où X_i est le nombre d'appels reçus durant la i^{e} période d'une minute, pour $i = 1, \ldots, 100$.

Question n$^{\text{o}}$ 2

Soit

$$f_{X,Y}(x, y) = \begin{cases} kx \text{ si } 0 < x < 1, 0 < y < x \\ 0 \text{ ailleurs} \end{cases}$$

la fonction de densité conjointe du vecteur aléatoire (X, Y).

(a) Trouver la constante k.

(b) Obtenir les fonctions de densité marginales de X et Y.

(c) Calculer VAR$[X]$ et VAR$[Y]$.

(d) Calculer le coefficient de corrélation de X et Y.

Question n$^{\text{o}}$ 3

Dans une banque, un guichet automatique permet de retirer des billets de 50 \$ ou 100 \$ à l'aide d'une carte magnétique. Il se peut aussi qu'un client ne puisse retirer d'argent si son compte n'est pas approvisionné ou s'il a fait une erreur de manipulation. Le nombre X de clients utilisant l'appareil dans un intervalle de cinq minutes est une variable aléatoire dont la fonction de probabilité $p_X(x)$ est

x	0	1	2
$p_X(x)$	0,3	0,5	0,2

De plus, on a constaté que le montant total Y retiré en cinq minutes est une variable aléatoire dont la fonction de probabilité conditionnelle $p_Y(y \mid X = x)$ est donnée par

y	0	50	100	150	200
$p_Y(y \mid X = 0)$	1	0	0	0	0
$p_Y(y \mid X = 1)$	0,1	0,7	0,2	0	0
$p_Y(y \mid X = 2)$	0,01	0,14	0,53	0,28	0,04

(a) Les variables aléatoires X et Y sont-elles indépendantes? Justifier votre réponse.

(b) Calculer la probabilité $P[X = 1, Y = 100]$.

(c) Calculer la probabilité $P[Y = 0]$.

(d) Calculer le nombre moyen de clients utilisant l'appareil en une heure.

Question n$^{\text{o}}$ 4

Un club privé décide de mettre sur pied un casino d'un soir dont les profits seront versés à une œuvre de charité. Les organisateurs ont décidé

- de demander à leurs membres de supporter les coûts fixes;
- d'admettre 1000 joueurs seulement, chacun avec une même mise initiale θ (en milliers de dollars) au début de la soirée;
- de choisir des jeux tels que le gain brut X_i (en milliers de dollars) du i^{e} joueur soit distribué uniformément sur l'intervalle $\left(0, \frac{3}{2}\theta\right)$.

On sait que la moyenne de la distribution $\text{U}\left(0, \frac{3}{2}\theta\right)$ est $\frac{3}{4}\theta$ et que sa variance est $\frac{3}{16}\theta^2$.

(a) Soit Y le gain brut total des 1000 joueurs. Donner la distribution approximative de Y ainsi que ses paramètres.

(b) Déterminer le montant θ que doit payer chaque joueur afin que le profit net (en milliers de dollars) du casino soit supérieur à 50 avec une probabilité de 0,95.

Question n$^{\text{o}}$ 5

Une voie rapide de circulation automobile a trois voies d'accès: A, B et C (voir la figure 4.5).

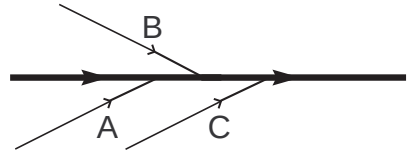

Fig. 4.5. Figure pour l'exercice n$^{\text{o}}$ 5

Le nombre de voitures accédant à la voie rapide durant une période d'une heure est défini par des variables aléatoires notées X_A, X_B et X_C, et ayant les caractéristiques suivantes:

	X_A	X_B	X_C
Moyenne	800	1000	600
Écart-type	40	50	30

Désignons par X le nombre total de voitures accédant à la voie rapide durant une période d'une heure.

(a) Calculer

(i) la moyenne de X;

(ii) l'écart-type de X, en supposant que les variables aléatoires X_A, X_B et X_C sont deux à deux indépendantes;

(iii) la probabilité que la variable aléatoire X soit comprise entre 2300 et 2500, si l'on suppose que les variables X_A, X_B et X_C sont indépendantes et distribuées approximativement normalement;

(iv) la probabilité que X soit supérieure à 2500 sous les mêmes hypothèses qu'en (iii).

(b) Soit Y le nombre de fois que X est supérieure ou égale à 2500 (sous les mêmes hypothèses que ci-desssus) durant 100 périodes (indépendantes) d'une heure.

(i) Donner la distribution de Y et ses paramètres.

(ii) Calculer, en utilisant une approximation basée sur la distribution normale, la probabilité que la variable aléatoire Y soit supérieure ou égale à 10.

(c) Calculer (i) la moyenne de X et (ii) son écart-type, si l'on suppose que les variables aléatoires X_A, X_B et X_C sont distribuées normalement et que les coefficients de corrélation sont: $\mathrm{CORR}[X_A, X_B] = 1/2$, $\mathrm{CORR}[X_A, X_C] = 4/5$ et $\mathrm{CORR}[X_B, X_C] = -1/2$.

Question n$^{\mathrm{o}}$ 6

La fonction de densité conjointe du couple de variables aléatoires (X, Y) est définie par (voir la figure 4.6):

$$f_{X,Y}(x,y) = \begin{cases} 3/4 \text{ si } -1 \le x \le 1,\ x^2 \le y < 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

(a) Calculer

(i) les fonctions de densité marginales de X et Y;

(ii) le coefficient de corrélation de X et Y.

(b) Les variables aléatoires X et Y sont-elles indépendantes? Justifier.

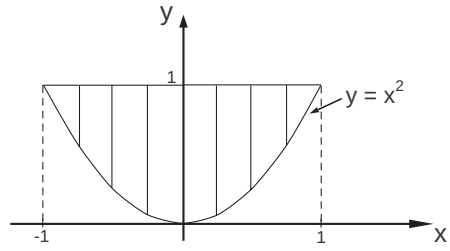

Fig. 4.6. Figure pour l'exercice n° 6

Question n° 7

Supposons que

$$f_{X,Y}(x,y) = \begin{cases} 4/\pi \text{ si } 0 < x \leq 1, 0 < y \leq \sqrt{2x - x^2} \\ 0 \quad \text{ailleurs} \end{cases}$$

Calculer la fonction de densité conditionnelle de Y étant donné que $X = x$.

Question n° 8

Calculer l'espérance mathématique de $(X + Y)^2$ si

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{8}(x+y) \text{ pour } 0 \leq x \leq 2, 0 \leq y \leq 2 \\ 0 \qquad \text{ailleurs} \end{cases}$$

Question n° 9

Supposons que X et Y sont deux variables aléatoires telles que $\text{VAR}[X] = \text{VAR}[Y] = 1$. On pose $Z = X - \frac{3Y}{4}$. Calculer le coefficient de corrélation de X et Z si $\text{COV}[X, Z] = \frac{1}{2}$.

Question n° 10

On lance une pièce de monnaie bien équilibrée jusqu'à ce que l'on obtienne "face", puis jusqu'à ce que l'on obtienne "pile". Si l'on suppose que les lancers sont indépendants, quelle est la probabilité que l'on doive lancer la pièce exactement neuf fois?

Question n° 11

Supposons que $X_1 \sim \text{N}(2,4)$, $X_2 \sim \text{N}(4,2)$ et $X_3 \sim \text{N}(4,4)$ sont des variables aléatoires indépendantes. Calculer le 75$^{\text{e}}$ centile de la variable aléatoire $Y := X_1 - 2X_2 + 4X_3$.

Question n° 12

La durée de vie d'un certain type de pneu présente (approximativement) une distribution normale ayant une moyenne de 25.000 km et un écart-type de 5000 km. On prend deux pneus au hasard (de façon indépendante). Quelle est la probabilité que l'un des deux pneus dure au moins 10.000 km de plus que l'autre?

Question n° 13

Une usine fabrique des articles dont le poids moyen est de 1,62 kg, avec un écart-type de 0,05 kg. Quelle est (approximativement) la probabilité que le poids total d'un groupe de 100 de ces articles soit compris entre 161,5 kg et 162,5 kg?

Question n° 14

La fonction de densité conjointe du couple de variables aléatoires (X, Y) est

$$f_{X,Y}(x,y) = \begin{cases} x+y \text{ si } 0 \le x \le 1, 0 \le y \le 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

On trouve que $E[X] = \frac{7}{12}$ et $\mathrm{VAR}[X] = \frac{11}{144}$. Calculer le coefficient de corrélation de X et Y.

Question n° 15

Soit $X \sim \mathrm{N}(2,1)$ et $Y \sim \mathrm{N}(4,4)$ deux variables aléatoires indépendantes. Calculer $P\left[|2X - Y| < \sqrt{8}\right]$.

Question n° 16

Soit X_i, $i = 1, 2, \ldots, 100$, des variables aléatoires indépendantes qui présentent toutes une distribution gamma de paramètres $\alpha = 9$ et $\lambda = 1/3$. Calculer approximativement $P\left[\bar{X} \ge 26\right]$, où $\bar{X} := \frac{1}{100}\sum_{i=1}^{100} X_i$.

Question n° 17

Soit X le nombre de boucles "do" dans un programme Fortran et soit Y le nombre d'essais nécessaires à un débutant pour faire fonctionner le programme. On suppose que la fonction de probabilité conjointe de (X, Y) est donnée par le tableau suivant:

$x \backslash y$	1	2	3
0	0,05	0,15	0,10
1	0,10	0,20	0,10
2	0,15	0,10	0,05

(a) Calculer $E[XY]$.

(b) Évaluer la probabilité $P[Y \ge 2 \mid X = 1]$.

(c) Les variables aléatoires X et Y sont-elles indépendantes? Justifier.

Question n$^{\text{o}}$ 18

Soit

$$f_{X,Y}(x,y) = \begin{cases} 6x \text{ si } 0 < x < 1, x < y < 1 \\ 0 \text{ ailleurs} \end{cases}$$

la fonction de densité conjointe de (X, Y).

(a) Calculer la fonction de densité marginale de X et celle de Y.

(b) Évaluer la probabilité $P[XY < 1/4]$.

Question n$^{\text{o}}$ 19

Un appareil est constitué de deux composants indépendants. L'un des composants est placé en attente et se met à fonctionner lorsque l'autre tombe en panne. La durée de vie (en heures) de chaque composant présente une distribution exponentielle de paramètre $\lambda = 1/2000$. Soit X la durée de vie de l'appareil.

(a) Donner la distribution de X et ses paramètres.

(b) Quelle est la moyenne de X?

Question n$^{\text{o}}$ 20

Le poids (en kilogrammes) d'objets manufacturés présente approximativement une distribution normale de paramètres $\mu = 1$ et $\sigma^2 = 0{,}02$. On prend 100 objets au hasard. Soit X_j le poids du j^{e} objet, pour $j = 1, 2, \ldots, 100$. On suppose que les X_j sont des variables aléatoires indépendantes.

(a) Calculer $P[X_1 - X_2 < 0{,}05]$.

(b) Trouver le nombre b tel que $P[X_1 + X_2 < b] = 0{,}025$.

(c) Calculer approximativement, en utilisant une distribution normale, la probabilité qu'exactement 70 des 100 objets considérés aient un poids inférieur à 1,072 kg.

Question n$^{\text{o}}$ 21

Soit

$$f_{X,Y}(x,y) = \begin{cases} e^{-2x} \text{ si } x > 0, 0 < y < 2 \\ 0 \quad \text{ailleurs} \end{cases}$$

la fonction de densité conjointe du vecteur aléatoire (X, Y).

(a) Calculer $f_X(x)$ et $f_Y(y)$.

(b) Quel est le coefficient de corrélation de X et Y? Justifier.

(c) Quel est le 50$^{\text{e}}$ centile de Y?

(d) Calculer $P\left[Y < e^X\right]$.

Question n^o 22

Le temps T (en années) entre deux pannes majeures d'électricité dans une région particulière présente une distribution exponentielle de moyenne 1,5. La durée X (en heures) des pannes présente approximativement une distribution normale de moyenne 4 et d'écart-type 2. On suppose que les pannes se produisent indépendamment les unes des autres.

(a) Étant donné qu'il n'y a pas eu de panne majeure depuis un an, quelle est la probabilité qu'il ne s'en produise aucune durant les neuf prochains mois?

(b) Combien de temps, au plus, durent 95 % des pannes majeures?

(c) Calculer la probabilité que la durée de la prochaine panne majeure et celle de la suivante diffèrent d'au plus 30 minutes.

(d) Calculer la probabilité que la panne majeure la plus longue, parmi les trois prochaines, dure moins de cinq heures.

(e) Utiliser le théorème central limite pour calculer (approximativement) la probabilité que la 30^e panne majeure à survenir à partir de maintenant se produise d'ici 50 ans.

Question n^o 23

Supposons que

$$f_{X,Y}(x,y) = \begin{cases} 2 \text{ si } 0 \leq x \leq 1, 0 \leq y \leq x \\ 0 \text{ ailleurs} \end{cases}$$

est la fonction de densité conjointe du vecteur aléatoire (X, Y).

(a) Calculer les fonctions de densité marginales de X et Y.

(b) Calculer $P[X - Y < 1/2]$.

(c) X et Y sont-elles indépendantes? Justifier.

(d) Calculer $E[XY]$.

Question n^o 24

Soit X le nombre de clients qui se présentent à un vendeur d'automobiles pendant une journée. On suppose que X présente une distribution de Poisson de paramètre $\lambda = 3$. De plus, on suppose qu'un client sur cinq, en moyenne, achète une voiture (lors d'une visite donnée), et ce, indépendamment des autres clients. Soit Y le nombre de voitures vendues par le vendeur en une journée.

(a) Étant donné que le vendeur a eu cinq clients pendant une journée donnée, quelle est la probabilité qu'il ait vendu exactement deux voitures?

(b) Quel est le nombre moyen de voitures vendues par le vendeur en une journée? Justifier.

Indication. On a: $E[Y] = \sum_{x=0}^{\infty} E[Y \mid X = x]\, P[X = x]$. De plus, sachant que $X = x$, Y est une variable aléatoire binomiale.

(c) Quelle est la probabilité que le vendeur ne vende aucune voiture pendant une journée donnée?

Indication. On a: $\sum_{x=0}^{\infty} \frac{k^x}{x!} = e^k$.

(d) Sachant que le vendeur n'a vendu aucune voiture au cours d'une journée donnée, quelle est la probabilité qu'il n'ait eu aucun client?

Question n$^{\text{o}}$ 25

Soit $X_1, X_2, \ldots, X_{50}$ des variables aléatoires indépendantes qui présentent toutes une distribution exponentielle de paramètre $\lambda = 2$.

(a) Calculer $P\left[X_1^2 > 4 \mid X_1 > 1\right]$.

(b) Soit $S = \sum_{i=1}^{50} X_i$.

(i) Donner la distribution de probabilité exacte de S ainsi que son ou ses paramètres.

(ii) Calculer approximativement, en utilisant le théorème central limite, la probabilité $P[S < 24]$.

Question n$^{\text{o}}$ 26

Soit $X_1 \sim \mathrm{N}(0, 1)$ et $X_2 \sim \mathrm{N}(1, 3)$ deux variables aléatoires indépendantes.

(a) Calculer $P[|X_1 - X_2| > 1]$.

(b) Quel est le 90$^{\text{e}}$ centile de $Y := X_1 + X_2$?

Question n$^{\text{o}}$ 27

La fonction de densité conjointe du vecteur aléatoire (X, Y, Z) est donnée par

$$f_{X,Y,Z}(x, y, z) = \begin{cases} k\left(\frac{x}{y} + z\right) & \text{si } 0 < x < 1,\ 1 < y < e,\ -1 < z < 1 \\ 0 & \text{ailleurs} \end{cases}$$

(a) Trouver la constante k.

(b) Montrer que

$$f_{X,Y}(x, y) = \begin{cases} 2x/y & \text{si } 0 < x < 1,\ 1 < y < e \\ 0 & \text{ailleurs} \end{cases}$$

(c) X et Y sont-elles des variables aléatoires indépendantes? Justifier.

(d) Calculer l'espérance mathématique de Y/X.

Question n^{o} 28

Soit X une variable aléatoire qui présente une distribution gamma de paramètres $\alpha = 25$ et $\lambda = 1/2$.

(a) Calculer la probabilité $P[X < 40]$ en utilisant une distribution de Poisson.
(b) Utiliser le théorème central limite pour calculer (approximativement) la probabilité $P[40 \leq X \leq 50]$. Justifier l'utilisation du théorème central limite.

Question n^{o} 29

Soit X une variable aléatoire qui présente une distribution binomiale de paramètres $n = 100$ et $p = 1/2$, et soit Y une variable aléatoire qui présente une distribution normale de paramètres $\mu = 50$ et $\sigma^2 = 25$.

(a) Calculer approximativement $P[X < 40]$.

(b) Calculer la probabilité $P[X = Y]$.

(c) Quel est le 33^{e} centile de Y?

Question n^{o} 30

Le tableau ci-dessous présente la distribution de probabilité conjointe du vecteur aléatoire (X, Y):

$x \setminus y$	1	2	3
0	1/9	2/9	1/9
1	1/18	1/9	1/18
2	1/6	1/18	1/9

Calculer $E[2XY]$.

Question n^{o} 31

Soit

$$f_{X,Y}(x, y) = \begin{cases} 4xy \text{ si } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

la fonction de densité conjointe du vecteur aléatoire (X, Y). Calculer la probabilité $P\left[X^2 + Y^2 < 1/4\right]$.

Question n^{o} 32

Soit X_1, X_2 et X_3 des variables aléatoires telles que (i) X_1 et X_2 sont indépendantes, (ii) $\text{VAR}[X_i] = 2$ pour $i = 1, 2, 3$, (iii) $\text{COV}[X_1, X_3] = 1/2$ et (iv) $\text{COV}[X_2, X_3] = 1$. Calculer $\text{VAR}[X_1 + X_2 + 2X_3]$.

Question n° 33

La durée de vie (en années) d'un appareil présente approximativement une distribution N$(5, 4)$. Utiliser le théorème central limite pour calculer la probabilité qu'au plus 10, parmi 30 appareils (indépendants) de ce type, durent au moins 6 ans.

Question n° 34

La fonction de densité conjointe du vecteur aléatoire (X, Y) est donnée par

$$f_{X,Y}(x, y) = \begin{cases} 1/\pi \text{ si } x^2 + y^2 \leq 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

(a) Vérifier que $f_{X,Y}$ est bien une fonction de densité conjointe.

(b) Calculer les fonctions de densité marginales f_X et f_Y.

(c) X et Y sont-elles indépendantes? Justifier.

(d) Calculer $P\left[X^2 + Y^2 \geq 1/4\right]$.

Question n° 35

On effectue 30 lancers indépendants d'un dé bien équilibré. Soit X le nombre de "6" obtenus et Y la somme de tous les nombres obtenus.

(a) Utiliser une distribution de Poisson pour calculer (approximativement) $P[X > 5]$ (même si la probabilité de succès est assez grande).

(b) Utiliser le théorème central limite pour calculer approximativement la probabilité $P[100 \leq Y < 111]$.

Indication. Si W est le résultat du lancer d'un dé bien équilibré, alors on a: $E[W] = 7/2$ et $\text{VAR}[W] = 35/12$.

Question n° 36

Soit

$$f_X(x) = \begin{cases} 1/10 \text{ si } 0 \leq x \leq 10 \\ 0 \quad \text{ailleurs} \end{cases}$$

et

$$p_Y(y) = \begin{cases} 1/10 \text{ si } y = 1, 2, \ldots, 10 \\ 0 \quad \text{autrement} \end{cases}$$

On suppose que X et Y sont des variables aléatoires indépendantes. Calculer (a) $P[X > Y]$; (b) $\text{VAR}[XY]$.

Question n^{o} 37

Supposons que X et Y sont deux variables aléatoires telles que $\mathrm{VAR}[X] = \mathrm{VAR}[Y] = 1$ et $\mathrm{COV}[X, Y] = 2/3$. Pour quelle valeur de k le coefficient de corrélation de X et $Z := X + kY$ est-il égal à $2/3$?

Question n^{o} 38

Un appareil électronique est constitué de 10 composants, dont la durée de vie (en mois) présente une distribution exponentielle de moyenne 50. On suppose que les composants fonctionnent indépendamment les uns des autres. Soit T la durée de vie de l'appareil. Obtenir la fonction de densité de T si

(a) les composants sont placés en série;

(b) les composants sont placés en parallèle;

(c) les composants sont placés en attente; c'est-à-dire qu'un seul composant fonctionne à la fois et, lorsqu'il tombe en panne, qu'un autre composant (s'il en reste au moins un qui n'est pas en panne) prend immédiatement la relève.

Question n^{o} 39

Des ampoules électriques achetées pour éclairer une patinoire extérieure ont une durée de vie moyenne de 3000 heures, avec un écart-type de 339 heures, indépendamment d'une ampoule à l'autre. On suppose que la durée de vie des ampoules présente approximativement une distribution normale.

(a) S'il est plus économique de remplacer toutes les ampoules quand 20 % d'entre elles sont brûlées, plutôt que de changer les ampoules au besoin, après combien d'heures devrait-on les remplacer?

(b) Supposons que seules les ampoules *brûlées* ont été remplacées après t_1 heures, alors que 20 % des ampoules devraient être brûlées. Trouver le pourcentage des ampoules qui seront brûlées après une période de $\frac{1}{2}t_1$ heures supplémentaires.

Question n^{o} 40

Le vecteur aléatoire continu (X, Y) possède la fonction de densité conjointe suivante:

$$f_{X,Y}(x, y) = \begin{cases} 6(1 - x - y) \text{ si } x > 0,\ 0 < y < 1 - x \\ 0 \qquad\qquad \text{ailleurs} \end{cases}$$

(a) Calculer la fonction de densité marginale de Y.

(b) Calculer le 40^{e} centile de Y.

(c) Soit $Z = Y^3$. Calculer la fonction de densité de Z.

(d) Calculer la probabilité $P[X < Y]$.

Question n° 41

On lance un dé bien équilibré deux fois, de façon indépendante. Soit X le nombre de "5" et Y le nombre de "6" obtenus. Calculer

(a) la fonction de probabilité conjointe $p_{X,Y}(x, y)$, où $0 \leq x + y \leq 2$;

(b) la fonction $F_{X,Y}\left(\frac{1}{2}, \frac{3}{2}\right)$;

(c) l'écart-type de 2^X;

(d) le coefficient de corrélation de X et Y.

Question n° 42

Soit

x_1	-2	-1	1	2
$p_{X_1}(x_1)$	1/3	1/6	1/3	1/6

et

x_2	0	1
$p_{X_2}(x_2 \mid X_1 = -2)$	1/2	1/2
$p_{X_2}(x_2 \mid X_1 = -1)$	1/2	1/2
$p_{X_2}(x_2 \mid X_1 = 1)$	1	0
$p_{X_2}(x_2 \mid X_1 = 2)$	0	1

(a) (i) Calculer la fonction de probabilité marginale de X_2.

(ii) Calculer $p_{X_2}\left(x_2 \mid \{X_1 = -2\} \cup \{X_1 = 2\}\right)$ pour $x_2 = 0$ et 1.

(b) Soit $Y = 2\,X_1 + X_1^2$. Calculer la fonction de répartition de Y.

Question n° 43

La durée X (en heures) des pannes majeures d'un certain métro présente approximativement une distribution normale de moyenne $\mu = 2$ et d'écart-type $\sigma = 0{,}75$. On suppose que les durées des pannes sont des variables aléatoires indépendantes.

(a) Calculer (de façon exacte) la probabilité que la durée de chacune de plus de 40 des 50 prochaines pannes majeures soit inférieure à 3 heures.

Remarque. Cette question requiert l'utilisation d'un logiciel ou d'une calculatrice.

(b) Utiliser une distribution normale pour calculer approximativement la probabilité en (a).

Question n° 44

Supposons que $X_1, \ldots, X_9$ sont des variables aléatoires indépendantes qui présentent toutes une distribution exponentielle de paramètre $\lambda = 1/2$.

(a) Calculer la probabilité

$$P\left[\frac{X_1 + X_2 + X_3}{X_4 + \ldots + X_9} < 1{,}5\right]$$

(b) (i) Soit $Y = X_1 + X_2$. Calculer $P[Y \leq y \mid X_1 = x_1]$, où $x_1 > 0$ et $y > 0$.

(ii) Obtenir la fonction de densité conditionnelle $f_Y(y \mid X_1 = x_1)$.

Question n^{o} 45

On considère une variable aléatoire discrète X qui présente une distribution hypergéométrique de paramètres $N = 10$, $n = 5$ et $d = 2$.

(a) On définit $Y = X^2$. Calculer le coefficient de corrélation de X et Y.

(b) Soit $X_1, X_2, \ldots, X_8$ des variables aléatoires indépendantes qui présentent la même distribution que X. On définit $Z = X_1 + \ldots + X_8$. Calculer la probabilité $P[8 \leq Z \leq 12]$.

Remarque. Cette question requiert l'utilisation d'un logiciel.

Question n^{o} 46

Supposons que la fonction de probabilité conjointe du vecteur aléatoire (X, Y) est donnée par

$$p_{X,Y}(x, y) = \begin{cases} \binom{x}{y} \dfrac{e^{-1}(1/2)^x}{x!} & \text{si } x = 0, 1, 2, \ldots;\ y = 0, 1, \ldots, x \\ 0 & \text{autrement} \end{cases}$$

(a) Obtenir les fonctions $p_X(x)$, $p_Y(y)$ et $p_Y(y \mid X = 35)$.

(b) Calculer la probabilité $P[12 < Y \leq 18 \mid X = 35]$

(i) de façon exacte (avec un logiciel, si possible);

(ii) en utilisant une approximation basée sur la distribution normale.

(c) Calculer la probabilité $P[X \leq 2 \mid X \geq 2]$.

Question n^{o} 47

Un système est constitué de trois composants, C_1, C_2 et C_3, placés en parallèle. La durée de vie T_1 (en années) du composant C_1 présente (approximativement) une distribution normale de paramètres $\mu = 4$ et $\sigma^2 = 2{,}25$. Dans le cas du composant C_2, sa durée de vie T_2 présente une distribution exponentielle de paramètre $\lambda = 1/4$. Finalement, la durée de vie T_3 du composant C_3 présente une distribution gamma de paramètres $\alpha = 2$ et $\lambda = 1/2$. De plus, on suppose que les variables aléatoires T_1, T_2 et T_3 sont indépendantes.

(a) Calculer la probabilité que le système fonctionne au bout d'une année.

(b) On considère 500 systèmes semblables à celui décrit ci-dessus. Calculer, en supposant que ces 500 systèmes sont indépendants, la probabilité que de 2 à 5 d'entre eux soient en panne au bout d'une année

(i) de façon exacte (à l'aide d'un logiciel ou d'une calculatrice);

(ii) en utilisant une approximation basée sur la distribution de Poisson.

(c) Supposons qu'au début seuls les composants C_1 et C_2 fonctionnent. Le composant C_3 est en attente et se met à fonctionner dès que C_1 et C_2 sont tous les deux en panne, ou au bout d'un an si C_1 ou C_2 fonctionne encore à ce moment. Supposons aussi que si C_1 ou C_2 fonctionne encore au bout d'un an, alors $T_3 \sim \mathrm{G}\big(2, \frac{1}{2}\big)$; sinon T_3 présente une distribution exponentielle de paramètre $\lambda = 1$. Calculer la probabilité que le composant C_3 fonctionne pendant au moins deux ans.

Question n° 48

Supposons que

$$f_{X,Y}(x,y) = \begin{cases} 1/8 \text{ si } x \geq 0,\ y \geq 0,\ 0 \leq x+y \leq 4 \\ 0 \quad \text{ailleurs} \end{cases}$$

est la fonction de densité conjointe du vecteur aléatoire (X, Y).

(a) Obtenir la fonction de densité marginale $f_X(x)$.

(b) Calculer (i) l'espérance mathématique de X, (ii) sa variance, (iii) son coefficient d'asymétrie β_1 et (iv) son coefficient d'aplatissement β_2.

(c) Calculer le coefficient de corrélation de X et Y. Les variables aléatoires X et Y sont-elles indépendantes? Justifier.

Question n° 49

Dans une certaine région, la température X (en °C) au cours du mois de septembre présente une distribution normale de paramètres $\mu = 15$ et $\sigma^2 = 25$. Calculer, en utilisant une approximation basée sur la distribution normale, la probabilité que la température excède 17 °C exactement 10 fois au cours de ce mois.

Question n° 50

Considérons la fonction de densité conjointe

$$f_{X,Y}(x,y) = \begin{cases} 90\,x^2 y(1-y) \text{ si } 0 < y < 1, 0 < x < y \\ 0 \qquad\qquad \text{ailleurs} \end{cases}$$

(a) Calculer les fonctions de densité marginales $f_X(x)$ et $f_Y(y)$.

(b) X et Y sont-elles des variables aléatoires indépendantes? Justifier.

(c) Calculer la covariance de X et Y.

Question n° 51

Soit X_1 et X_2 deux variables aléatoires discrètes dont la fonction de probabilité conjointe est donnée par

$$p_{X_1,X_2}(x_1,x_2) = \frac{2}{x_1!x_2!(2-x_1-x_2)!}\left(\frac{1}{4}\right)^{x_1}\left(\frac{1}{3}\right)^{x_2}\left(\frac{5}{12}\right)^{2-x_1-x_2}$$

si $x_1 \in \{0,1,2\}$, $x_2 \in \{0,1,2\}$ et $x_1 + x_2 \leq 2$, et $p_{X_1,X_2}(x_1,x_2) = 0$ autrement.

(a) Soit $Y_1 = X_1 + X_2$ et $Y_2 = X_1 - X_2$. Calculer les fonctions de probabilité de Y_1 et Y_2.

(b) Calculer la fonction $p_{Y_2}(y_2 \mid Y_1 = 2)$.

Question n° 52

On prend un nombre X au hasard sur l'intervalle $[-1,1]$ et un nombre Y au hasard sur l'intervalle $[-1,X]$.

(a) Calculer $f_{X,Y}(x,y)$ et $f_Y(y)$.

(b) Calculer (i) $E[(X+1)Y]$ et (ii) $E[Y]$.

(c) Utiliser la partie (b) pour calculer $\mathrm{COV}[X,Y]$.

Question n° 53

Le réservoir d'une station-service est normalement rempli d'essence tous les lundis. La capacité du réservoir est de 20.000 litres. On nous prévient, après avoir rempli le réservoir, que l'on ne pourra pas venir le remplir de nouveau le lundi suivant. Quelle est la probabilité que l'on ne puisse pas satisfaire la demande pendant deux semaines, si la demande hebdomadaire (en milliers de litres) présente une distribution

(a) exponentielle de paramètre $\lambda = 1/10$?

(b) gamma de paramètres $\alpha = 5$ et $\lambda = 1/2$?

Remarque. Utiliser un logiciel pour répondre à la question (b), si possible.

Question n° 54

Une variable aléatoire X possède la fonction de probabilité suivante:

x	-1	0	1
$p_X(x)$	1/8	3/4	1/8

Soit X_1 et X_2 deux variables aléatoires indépendantes et distribuées comme X. On pose $Y = X_2 - X_1$.

(a) Obtenir la fonction de probabilité conjointe du couple (X_1, X_2).

(b) Calculer le coefficient de corrélation de X_1 et Y.

(c) Les variables aléatoires X_1 et Y sont-elles indépendantes? Justifier.

Question n$^{\text{o}}$ 55

Soit $X_1, \ldots, X_{10}$ des variables aléatoires indépendantes qui présentent toutes une distribution exponentielle de paramètre $\lambda = 1$. On définit $Y = \sum_{i=1}^{10} X_i$.

(a) Évaluer, *sans* utiliser le théorème central limite, la probabilité suivante: $P[Y < 5]$.

(b) Utiliser le théorème central limite pour évaluer $P[Y \geq 10]$.

Questions à choix multiple

Question n$^{\text{o}}$ 1

Soit $X \sim \mathrm{N}(0,1)$ et $Y \sim \mathrm{N}(1,4)$ deux variables aléatoires telles que $\mathrm{COV}[X,Y] = 1$. Calculer $P[X + Y < 12]$.

(a) 0 (b) 0,6915 (c) 0,8413 (d) 0,9773 (e) 1

Question n$^{\text{o}}$ 2

Calculer $P[3 \leq X + Y < 6]$ si $X \sim \mathrm{Poi}(\lambda = 1)$ et $Y \sim \mathrm{Poi}\ (\lambda = 2)$ sont des variables aléatoires indépendantes.

(a) 0,269 (b) 0,493 (c) 0,543 (d) 0,726 (e) 0,916

Question n$^{\text{o}}$ 3

Utiliser l'approximation de la distribution binomiale par la distribution normale pour calculer $P[X \leq 12]$, où $X \sim \mathrm{B}(n = 25, p = 1/2)$.

(a) 0,4207 (b) 0,4681 (c) 0,5 (d) 0,5319 (e) 0,5793

Question n$^{\text{o}}$ 4

Soit

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{4\pi} \text{ si } x^2 + y^2 \leq 4 \\ 0 \text{ ailleurs} \end{cases}$$

Calculer $f_X(x)$.

(a) $\frac{1}{2\pi}\sqrt{4 - x^2}$ si $-2 \leq x \leq 2$ (b) $\frac{1}{2\pi}\sqrt{4 - x^2}$ si $0 \leq x \leq 2$
(c) $\frac{1}{4\pi}\sqrt{4 - x^2}$ si $-2 \leq x \leq 2$ (d) $\frac{1}{4\pi}\sqrt{4 - x^2}$ si $0 \leq x \leq 2$

(e) $\frac{1}{4\pi}\sqrt{4-x^2}$ si $-2 \leq x \leq y$

Question n° 5

Supposons que $p_X(x \mid Y = y) = 1/3$ pour $x = 0, 1, 2$ et $y = 1, 2$, et que $p_Y(y) = 1/2$ pour $y = 1, 2$. Calculer $p_{X,Y}(1, 2)$.

(a) $\frac{1}{6}$ (b) $\frac{1}{3}$ (c) $\frac{1}{2}$ (d) $\frac{2}{3}$ (e) 1

Question n° 6

Soit X une variable aléatoire telle que $E[X^n] = 1/2$ pour $n = 1, 2, \ldots$ On pose $Y = X^2$. Calculer $\rho_{X,Y}$.

(a) 0 (b) $\frac{1}{4}$ (c) $\frac{1}{2}$ (d) $\frac{3}{4}$ (e) 1

Question n° 7

Supposons que X est une variable aléatoire telle que $E[X] = 1$ et $\mathrm{VAR}[X] = 1$. Calculer approximativement la probabilité $P\left[\sum_{i=1}^{49} X_i < 56\right]$, où $X_1, X_2, \ldots, X_{49}$ sont des variables aléatoires indépendantes distribuées comme X.

(a) 0,5 (b) 0,6554 (c) 0,8413 (d) 0,8643 (e) 1

Question n° 8

On définit $W = 3X + 2Y - Z$, où X, Y et Z sont des variables aléatoires indépendantes telles que $\sigma_X^2 = 1$, $\sigma_Y^2 = 4$ et $\sigma_Z^2 = 9$. Calculer σ_W.

(a) $\sqrt{2}$ (b) 4 (c) $\sqrt{20}$ (d) $\sqrt{34}$ e) 10

Question n° 9

On considère la fonction de densité conjointe

$$f_{X,Y}(x, y) = \begin{cases} 1/4 \text{ si } 0 < x < 2, 0 < y < 2 \\ 0 \quad \text{ailleurs} \end{cases}$$

Calculer $P[X > 2Y]$.

(a) 1/8 (b) 1/4 (c) 1/2 (d) 3/4 (e) 7/8

Question n° 10

Calculer $P[2X < Y]$ si

$$f_{X,Y}(x, y) = \begin{cases} 2 \text{ pour } 0 \leq x \leq y \leq 1 \\ 0 \text{ ailleurs} \end{cases}$$

(a) 0 (b) 1/4 (c) 1/2 (d) 3/4 (e) 1

Question n° 11

On définit $X = \max\{X_1, X_2\}$, où X_1 et X_2 sont les nombres obtenus en lançant deux dés bien équilibrés simultanément. C'est-à-dire que X est le plus grand des deux nombres observés. Calculer $E[X]$.

(a) 91/36 (b) 3,5 (c) 4 (d) 161/36 (e) 4,5

Question n° 12

Supposons que

$$f_X(x \mid Y = y) = \begin{cases} \frac{1}{2y} \text{ si } 0 < x < 2y \\ 0 \text{ ailleurs} \end{cases}$$

et

$$f_Y(y) = \begin{cases} \frac{1}{2} \text{ si } 0 < y < 2 \\ 0 \text{ ailleurs} \end{cases}$$

Calculer $f_{X,Y}(x, y)$.

(a) $\frac{1}{8}$ si $0 < x < 4$ et $0 < y < 2$ (b) $\frac{1}{4y}$ si $0 < x < 4$ et $0 < y < 2$
(c) $\frac{1}{4y}$ si $0 < x < 2y$ et $0 < y < 2$ (d) $\frac{1}{2y}$ si $0 < x < 2y$ et $0 < y < 2$
(e) $\frac{1}{y}$ si $0 < x < 4$ et $0 < y < 2$

Question n° 13

Soit

x	-2	0	2
$p_X(x)$	1/8	3/4	1/8

On définit $Y = -X^2$. Calculer le coefficient de corrélation de X et Y.

(a) -1 (b) $-1/2$ (c) 0 (d) 1/2 (e) 1

Question n° 14

Supposons que X et Y sont deux variables aléatoires indépendantes. On définit deux autres variables aléatoires par $R = aX + b$ et $S = cY + d$. Pour quelles valeurs de a, b, c et d les variables R et S sont-elles non corrélées (c'est-à-dire que $\rho_{R,S} = 0$)?

(a) aucune (b) $a = b = 1$ (c) $b = d = 0$ (d) $a = c = 1$, $b = d = 0$
(e) toutes

Question n° 15

Supposons que X_1 et X_2 sont deux variables aléatoires indépendantes distribuées uniformément sur l'intervalle [0,1]. Soit X la plus petite des deux variables aléatoires. Calculer $P[X > 1/4]$.

(a) 1/16 (b) 1/8 (c) 1/4 (d) 9/16 (e) 3/4

Question n° 16

Calculer $P[X_1 + X_2 < 2]$ si X_1 et X_2 sont deux variables aléatoires indépendantes qui présentent une distribution exponentielle de paramètre $\lambda = 1$.

(a) 0,324 (b) 0,405 (c) 0,594 (d) 0,676 (e) 0,865

Question n° 17

Soit $X_1, \ldots, X_{36}$ des variables aléatoires indépendantes, où X_i présente une distribution gamma de paramètres $\alpha = 2$ et $\lambda = 3$. Calculer approximativement $P\left[\frac{2}{3} < \bar{X} < \frac{3}{4}\right]$, où $\bar{X} := \frac{1}{36}\sum_{i=1}^{36} X_i$.

(a) 0,218 (b) 0,355 (c) 0,360 (d) 0,497 (e) 0,855

Question n° 18

Supposons que $X_1, \ldots, X_n$ sont des variables aléatoires $\mathrm{N}(0,1)$ indépendantes. Quelle est la plus petite valeur de n pour laquelle on peut écrire que $P[-0{,}1n < \sum_{i=1}^{n} X_i < 0{,}1n] \geq 0,95$?

(a) 19 (b) 20 (c) 271 (d) 384 (e) 385

Question n° 19

Soit $X_1 \sim \mathrm{N}(0,1)$, $X_2 \sim \mathrm{N}(-1,1)$ et $X_3 \sim \mathrm{N}(1,1)$ des variables aléatoires indépendantes. Calculer $P[|X_1 + 2X_2 - 3X_3| > 5]$.

(a) 0,004 (b) 0,496 (c) 0,5 (d) 0,504 (e) 0,996

Question n° 20

Calculer approximativement, à l'aide d'une distribution normale, la probabilité $P\left[\sum_{i=1}^{100} X_i \leq 251\right]$, où $X_1, \ldots, X_{100}$ sont des variables aléatoires indépendantes telles que X_i présente une distribution binomiale de paramètres $n = 10$ et $p = 1/4$ pour $i = 1, \ldots, 100$.

(a) 0,50 (b) 0,51 (c) 0,53 (d) 0,56 (e) 0,59

5

Statistique descriptive et estimation

Avec ce chapitre, nous commençons l'étude de la statistique à proprement parler. D'abord, nous allons voir les principaux concepts de ce que l'on appelle *statistique descriptive*. Cette matière se trouve souvent avant la partie probabilités dans les manuels de probabilités et statistique pour scientifiques et ingénieurs, car elle ne fait pas appel aux notions de probabilités. Toutefois, il nous semble préférable de présenter la statistique descriptive immédiatement avant le début de la *statistique mathématique* de base. Ainsi, il est plus facile de bien faire la distinction entre des quantités calculées en utilisant des données que l'on a recueillies et les quantités théoriques correspondantes.

Dans ce chapitre, nous verrons comment *estimer* les paramètres inconnus qui apparaissent dans les fonctions de probabilité ou les fonctions de densité des variables aléatoires, en particulier les paramètres des différents modèles que nous avons étudiés au chapitre 3, par exemple les distributions de Poisson, exponentielle, normale, etc. Nous allons aussi présenter les *distributions d'échantillonnage* les plus importantes, lesquelles sont des distributions particulières très utilisées en statistique. Enfin, nous terminerons ce chapitre par la notion d'*intervalles de confiance* pour les paramètres inconnus.

5.1 Statistique descriptive

Définition 5.1.1. *Un* **échantillon aléatoire** *de* **taille** n *d'une variable aléatoire X est un ensemble $X_1, \ldots, X_n$ de variables aléatoires indépendantes et dont la distribution est identique à celle de X. C'est-à-dire que X_k possède la même fonction de répartition que X, pour $k = 1, \ldots, n$. La variable aléatoire X est aussi appelée* **population** *et chaque X_k est une* **observation** *de X.*

Remarque. Une valeur prise par une observation X_k est un nombre réel x_k, appelé **donnée** ou **observation particulière** de X.

Dans cette section, nous allons présenter quelques outils qui permettent de visualiser un ensemble de données $x_1, \ldots, x_n$ qui ont été recueillies. Nous appellerons cet ensemble de données **échantillon** ou bien **échantillon aléatoire particulier** de X. Nous allons aussi définir des quantités qui peuvent être calculées avec ces données. Toutes ces quantités ont leur équivalent dans le cas d'observations $X_1, \ldots, X_n$ de la variable aléatoire X.

5.1.1 Tableaux d'effectifs ou de fréquences

Supposons que l'on s'intéresse à la durée de vie (en heures) des ampoules électriques d'une certaine marque. On prend 20 ampoules au hasard et on mesure leur durée de vie; on obtient:

115 2456 534 3915 1046 1916 1117 1303 865 340
575 3563 4413 500 2096 149 1511 2244 695 1021

En se servant de ces 20 données, on peut construire un **tableau d'effectifs** de la façon suivante: on prend un intervalle $[a, b]$ qui couvre l'ensemble des données et on le divise en k ($5 \leq k \leq 20$, en général) sous-intervalles, appelés **classes**, de même longueur, disjoints et exhaustifs. On associe à chaque classe le nombre de données qu'elle compte, appelé son **effectif**.

Remarques.
(i) Parfois, au lieu d'indiquer la classe, on donne uniquement son point milieu, ou bien la limite inférieure des classes.

(ii) On peut aussi construire un **tableau d'effectifs cumulatifs**, de **pourcentages** (en divisant chaque effectif par le nombre total de données) et de **pourcentages cumulatifs**.

(iii) Les limites des classes devraient avoir le même nombre de chiffres significatifs que les données elles-mêmes.

Exemple 5.1.1. La plus petite valeur parmi les 20 données ci-dessus est 115, et la plus grande 4413. Alors, si l'on divise l'intervalle $[0, 5000]$ en cinq classes de longueur 1000, on obtient le tableau suivant:

Classe	*Effectif*	*Effec. cumulatif*	*Pourcentage*	*Pourc. cumulatif*
[0,1000)	8	8	40	40
[1000,2000)	6	14	30	70
[2000,3000)	3	17	15	85
[3000,4000)	2	19	10	95
[4000,5000)	1	20	5	100

♢

Remarque. Le choix de la limite inférieure, a, et du nombre de classes a beaucoup d'importance, surtout si le nombre total de données est petit.

Une autre façon de représenter les données est au moyen d'un **diagramme de points** obtenu en plaçant chaque valeur différente recueillie sur l'axe des x et, au-dessus de celle-ci, un point chaque fois que l'on a obtenu cette valeur. Notons que, dans le cas des données ci-dessus, aucune valeur n'a été obtenue plus d'une fois.

Finalement, un **tableau "tige-et-feuille"**, aussi appelé **histogramme de Tukey**, est construit en divisant chaque donnée en deux parties; par exemple, le premier ou les deux premiers chiffres constituent la **tige** et le reste (ou le chiffre suivant) la **feuille**.

Exemple 5.1.1 (suite). Si l'on prend le premier chiffre (et en écrivant que 115 = 0115) comme tige, on obtient:

0 || 115, 149, 340, 500, 534, 575, 695, 865		0 || 11355568
1 || 021, 046, 117, 303, 511, 916		1 || 001359
2 || 096, 244, 456	ou	2 || 024
3 || 563, 915		3 || 59
4 || 413		4 || 4

♢

Remarque. En regardant (de côté) un tableau tige-et-feuille, on peut avoir une idée de la distribution de la variable aléatoire qui a généré les données.

5.1.2 Représentations graphiques

Définition 5.1.2. *Un* **polygone d'effectifs** *est obtenu en faisant le graphique des effectifs en fonction des points milieux des classes et en reliant les points du graphique par des segments de droite.*

Remarques.
(i) On rajoute deux classes (vides) aux extrémités pour fermer le polygone.
(ii) On a aussi des **polygones d'effectifs cumulatifs**, de **pourcentages** et de **pourcentages cumulatifs**.

Exemple 5.1.1 (suite). En utilisant les données de l'exemple 5.1.1, on construit le polygone d'effectifs de la figure 5.1. ♢

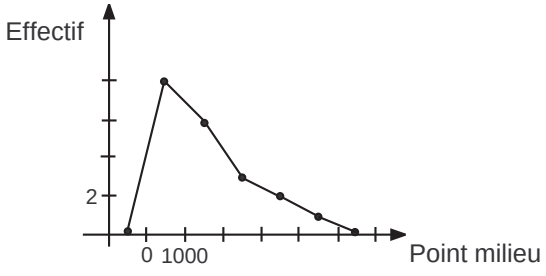

Fig. 5.1. Polygone d'effectifs construit en se servant des données de l'exemple 5.1.1

Définition 5.1.3. *Un* **histogramme**, *aussi appelé* **diagramme en barres**, *est obtenu en plaçant un rectangle au-dessus de chaque classe; la hauteur du rectangle est égale à l'effectif de la classe.*

Exemple 5.1.1 (suite). Avec les données de l'exemple 5.1.1, on obtient l'histogramme de la figure 5.2. ♢

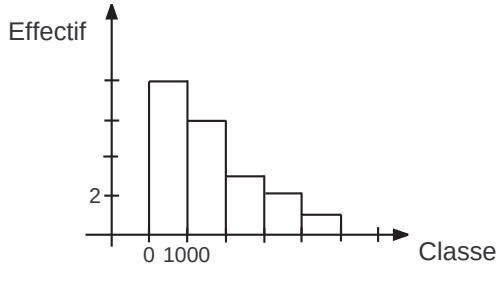

Fig. 5.2. Histogramme obtenu avec les données de l'exemple 5.1.1

5.1.3 Quantités calculées en utilisant les données

Définition 5.1.4. *Supposons que l'on dispose de n données:* $x_1, x_2, \ldots, x_n$. *La* **moyenne de l'échantillon** *est définie par*

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{5.1}$$

Remarques.
(i) La quantité $\bar{x}$ est la moyenne *arithmétique* des données. Dans le cas d'un échantillon aléatoire de X, la quantité équivalente $\bar{X}$ sert à *estimer* la moyenne de la *variable aléatoire* X qui a généré les observations $X_1, X_2, \ldots, X_n$.

(ii) On peut associer un *poids* p_i à chaque donnée, de façon que $p_i \geq 0$ pour tout i et que $\sum_{i=1}^{n} p_i = 1$; alors l'expression $\sum_{i=1}^{n} p_i x_i$ est appelée **moyenne pondérée** des x_i.

(iii) Lorsqu'on ne dispose que d'un tableau d'effectifs, on peut approcher la moyenne de l'échantillon en supposant que toutes les données dans une classe quelconque sont égales au point milieu de cette classe.

La moyenne d'un échantillon est une *mesure de position centrale*. On définit maintenant une *mesure de dispersion* des données autour de leur moyenne.

Définition 5.1.5. *La* **variance de l'échantillon** *est donnée par*

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \tag{5.2}$$

*De plus, l'***écart-type** *de l'échantillon est* $s = \sqrt{s^2}$.

Remarques.
(i) Certains auteurs définissent la variance de l'échantillon par $\sum_{i=1}^{n}(x_i - \bar{x})^2/n$. La raison pour laquelle on divise, en général, la somme des carrés par $n-1$ plutôt que par n est qu'en utilisant $\bar{X}$, dans la définition équivalente pour le cas d'un échantillon aléatoire de X, on perd un *degré de liberté*. En effet, soit $X_1, \ldots, X_n$ n observations indépendantes de X; si l'on nous donne $n-1$ observations et la moyenne (ou la somme) des n observations, alors on peut en déduire la valeur de la n^{e} observation.

(ii) Les calculatrices donnent généralement l'écart-type s des données plutôt que s^2. Souvent, elles donnent à la fois l'écart-type obtenu en divisant la somme des carrés par $n-1$, et celui obtenu en divisant cette somme par n. Dans ce cas, ces quantités peuvent être notées σ_{n-1} et σ_n, respectivement. On trouve aussi des calculatrices qui utilisent la notation s comme ici, et la notation σ pour l'écart-type calculé en divisant la somme des carrés par n.

Pour *estimer* le k^{e} moment par rapport à la moyenne (voir la page 108) de la variable aléatoire X, on utilise

$$\hat{\mu}_k := \sum_{i=1}^{n} \frac{(x_i - \bar{x})^k}{n-1} \tag{5.3}$$

Il s'ensuit que

$$\hat{\beta}_1 = \frac{\hat{\mu}_3^2}{s^6} \quad \text{et} \quad \hat{\beta}_2 = \frac{\hat{\mu}_4}{s^4} \tag{5.4}$$

Remarque. La notation "$\hat{}$" désigne ce que l'on appelle *estimations ponctuelles* (voir la section 5.2) des quantités théoriques correspondantes.

Notation. On écrira $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ pour indiquer que les données ont été placées en ordre croissant; c'est-à-dire que

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} \tag{5.5}$$

En utilisant cette notation, on peut définir les quantités suivantes.

Définition 5.1.6. *(i) L'***étendue** *(en anglais,* range*) des données est*

$$R = x_{(n)} - x_{(1)} \tag{5.6}$$

(ii) La **médiane** *des données est définie par*

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \textit{si } n \textit{ est impair} \\ \dfrac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \textit{si } n \textit{ est pair} \end{cases} \tag{5.7}$$

Remarques.
(i) L'étendue n'est pas très utile, sauf pour la construction d'un tableau d'effectifs. On peut aussi s'en servir en *contrôle de la qualité.*

(ii) La médiane d'un échantillon est *unique*. Cependant, il peut y avoir plus d'une donnée égale à la médiane. Par exemple, si l'on a: $x_1 = 1$, $x_2 = 3$, $x_3 = 2$ et $x_4 = 2$, alors on peut écrire que $x_3 = x_4 = \tilde{x} = 2$.

(iii) La médiane est aussi notée m ou Md.

Définition 5.1.7. *Le* **mode** *de l'échantillon est soit la donnée la plus fréquente, soit le point milieu de la classe ayant le plus grand effectif.*

Remarque. On définit aussi, entre autres quantités, le **coefficient de variation** de l'échantillon par

$$\mathrm{CV} = \frac{s}{\bar{x}} \cdot 100\ \% \tag{5.8}$$

Il s'agit donc d'une mesure de dispersion *relative* des données autour de la moyenne $\bar{x}$ de l'échantillon. De plus, l'**écart-type de la moyenne** est donné par $s/\sqrt{n}$.

Exemple 5.1.1 (suite et fin). Si l'on calcule les quantités définies ci-dessus avec les 20 données fournies au début de cette section, on trouve que

$\bar{x} = 1518{,}7$	$\hat{\beta}_1 \simeq 0{,}95$	$\tilde{x} = 1081{,}5$
$s^2 = 1.584.881$	$\hat{\beta}_2 \simeq 2{,}83$	mode $= 500$ (2^{e} définition)
$s \simeq 1258{,}92$	$R = 4298$	CV $\simeq 82{,}9\ \%$

♢

Maintenant, le but de la statistique descriptive est aussi de trouver un modèle théorique pour les données que l'on a recueillies; c'est-à-dire, dans l'exemple de cette section, de trouver la distribution que présente la variable aléatoire X qui désigne la durée de vie d'une ampoule quelconque. Puisque les 20 données ont en fait été obtenues en générant des observations particulières d'une variable aléatoire X qui présente une distribution $\mathrm{Exp}(\lambda = 0{,}001)$, on peut comparer les quantités calculées en utilisant les données aux valeurs théoriques:

$\mu = 1/\lambda = 1000$	$\beta_1 = 4$	$x_m \simeq 693$
$\sigma^2 = 1/\lambda^2 = 1.000.000$	$\beta_2 = 9$	mode $= 0$
$\sigma = 1000$	$R = \infty$	$\dfrac{\sigma}{\mu} \cdot 100\ \% = 100\ \%$

Remarques.

(i) Le nombre de données, $n = 20$, est beaucoup trop petit pour pouvoir espérer que les quantités calculées avec les données et les valeurs théoriques soient

presque égales. Cependant, l'allure du polygone d'effectifs ressemble bien à celle de la fonction de densité d'une distribution exponentielle. Avec $n = 200$, les résultats suivants ont été obtenus:

$$\begin{array}{lll} \bar{x} \simeq 1036{,}53 & \hat{\beta}_1 \simeq 2{,}70 & \tilde{x} = 711{,}87 \\ s^2 = 1.059.190 & \hat{\beta}_2 \simeq 6{,}08 & \text{mode} = 150 \\ s \simeq 1029{,}17 & R \simeq 5824{,}13 & \text{CV} \simeq 99{,}29\ \% \end{array}$$

Notons que le mode a été calculé à partir d'un tableau d'effectifs constitué de 20 classes de 0 à 6000. De plus, en générant 200 autres observations particulières d'une distribution $\text{Exp}(\lambda = 0{,}001)$, les résultats pourraient être sensiblement différents.

(ii) On verra au chapitre 6 comment *tester* l'hypothèse que les observations proviennent d'une distribution donnée; par exemple, ici on ne rejette *pas* (au *seuil de signification* $\alpha = 0{,}01$) l'hypothèse que les 20 nombres obtenus proviennent d'une distribution exponentielle de paramètre $\lambda = 0{,}001$ (mais, encore une fois, $n = 20$ est trop petit pour pouvoir tirer des conclusions assez fiables).

Lorsqu'on possède des observations particulières, $(x_1, y_1), \ldots, (x_n, y_n)$, d'un couple de variables aléatoires (X, Y), en plus des moyennes, $\bar{x}$ et $\bar{y}$, et des variances, s_x^2 et s_y^2, on définit la **covariance** des données par

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{5.9}$$

La covariance est une mesure de *dispersion conjointe* des données autour des moyennes $\bar{x}$ et $\bar{y}$.

Finalement, on définit aussi le **coefficient de corrélation** (de Pearson) des données comme suit:

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} \tag{5.10}$$

Cette quantité est une mesure de liaison *linéaire* entre les variables X et Y qui ont généré ces données. Il s'agit, en fait, de l'*estimation ponctuelle* du coefficient de corrélation théorique $\rho_{X,Y}$ défini au chapitre 4, à la page 163. Comme dans le cas de $\rho_{X,Y}$, on peut montrer que $-1 \leq r_{x,y} \leq 1$.

5.2 Estimation ponctuelle

Définition 5.2.1. *Une* **statistique** *est une fonction, qui ne dépend d'aucun paramètre inconnu, des observations dans un échantillon aléatoire. La distribu-*

tion que présente une statistique est appelée **distribution** *(ou* **loi***)* **d'échantillonnage**.

Exemple 5.2.1. La **moyenne de l'échantillon aléatoire**:

$$\bar{X} := \sum_{k=1}^{n} \frac{X_k}{n} \tag{5.11}$$

est une statistique. De même, la **variance de l'échantillon aléatoire**:

$$S^2 := \sum_{k=1}^{n} \frac{(X_k - \bar{X})^2}{n-1} \tag{5.12}$$

est une statistique. Cependant, la quantité

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

est une statistique seulement si $\mu := E[X]$ est connu. Finalement, si X présente une distribution normale, alors la distribution d'échantillonnage de $\bar{X}$ est normale; si X ne présente pas une distribution normale, mais si n est suffisamment grand, alors $\bar{X}$ présente approximativement une distribution normale, par le théorème central limite. ♢

Définition 5.2.2. *Une* **estimation ponctuelle** *d'un paramètre inconnu, θ, d'une population est un nombre qui correspond à ce paramètre.*

Définition 5.2.3. *Un* **estimateur** *d'un paramètre inconnu θ d'une population est une statistique $T = g(X_1, \ldots, X_n)$ qui correspond à ce paramètre.*

Remarques.
(i) Un estimateur T est une variable aléatoire, puisque c'est une fonction d'un échantillon aléatoire.
(ii) On désigne souvent un estimateur quelconque de θ par $\hat{\theta}$. Notons que $\hat{\theta}$ peut aussi désigner une estimation ponctuelle de θ, comme à la section précédente. Le contexte devrait permettre de savoir quelle interprétation est la bonne.
(iii) La quantité θ peut, en fait, être un vecteur: $\theta = (\theta_1, \ldots, \theta_m)$.

Exemple 5.2.2. La moyenne de l'échantillon aléatoire, $\bar{X}$, est un estimateur de la moyenne (inconnue) μ de la population. De même, S^2 est un estimateur

de $\sigma^2 := \mathrm{VAR}[X]$. De plus, des valeurs obtenues avec un échantillon aléatoire *particulier*, c'est-à-dire $\bar{x}$ et s^2, sont des estimations ponctuelles de μ et σ^2, respectivement. ♢

Remarque. Il est important de faire la distinction entre les trois moyennes et les trois variances définies dans ce chapitre, comme nous l'avons résumé dans le tableau suivant:

	Population	Échantillon aléatoire	Éch. al. particulier
Moyenne	μ	$\bar{X}$	$\bar{x}$
Variance	σ^2	S^2	s^2

Il faut se rappeler que μ et σ^2 sont des *constantes* (généralement) inconnues que l'on doit estimer à l'aide d'un échantillon de la population X. Les quantités $\bar{X}$ et S^2 sont des *variables aléatoires* qui servent à estimer μ et σ^2, respectivement. Enfin, $\bar{x}$ et s^2 sont des valeurs particulières prises par $\bar{X}$ et S^2, c'est-à-dire des *constantes* réelles que l'on peut calculer en se servant d'une calculatrice, par exemple. Parfois, on suppose, pour simplifier, que la variance σ^2 de la population est connue; dans ce cas, il ne sert à rien de chercher à l'estimer.

5.2.1 Propriétés des estimateurs

Définition 5.2.4. *Un estimateur T d'un paramètre inconnu θ est dit* **sans biais** *ou* **non biaisé** *si $E[T] = \theta$. De plus, le* **biais** *de T est défini par*

$$\mathrm{Biais}[T] = E[T] - \theta \tag{5.13}$$

Remarque. L'estimateur $T\,(= T(n))$ est dit *asymptotiquement sans biais* si

$$\lim_{n\to\infty} E[T] = \theta \tag{5.14}$$

Exemple 5.2.3. La moyenne $\bar{X}$ de l'échantillon aléatoire est un estimateur non biaisé de μ, car

$$E[\bar{X}] = E\left[\sum_{k=1}^{n} \frac{X_k}{n}\right] = \frac{1}{n}\sum_{k=1}^{n} E[X_k] \overset{\text{i.d.}}{=} \frac{1}{n}\sum_{k=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

Ce résultat s'écrit sous la forme

$$E[\bar{X}] = E[X] \tag{5.15}$$

De même, S^2 est un estimateur non biaisé de σ^2. En effet, on a:

$$S^2 := \sum_{k=1}^{n} \frac{(X_k - \bar{X})^2}{n-1} = \frac{1}{n-1}\left\{ \sum_{k=1}^{n} X_k^2 - 2\bar{X} \underbrace{\sum_{k=1}^{n} X_k}_{n\bar{X}} + n\bar{X}^2 \right\}$$

$$\Longrightarrow$$

$$S^2 = \frac{1}{n-1}\left[\sum_{k=1}^{n} X_k^2 - n\bar{X}^2 \right] \tag{5.16}$$

De plus (voir la section 4.4),

$$\mathrm{VAR}[\bar{X}] = \mathrm{VAR}\left[\sum_{k=1}^{n} \frac{X_k}{n} \right] \overset{\text{ind.}}{=} \sum_{k=1}^{n} \frac{\mathrm{VAR}[X_k]}{n^2} \overset{\text{i.d.}}{=} \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \tag{5.17}$$

C'est-à-dire que

$$\mathrm{VAR}[\bar{X}] = \frac{\mathrm{VAR}[X]}{n} \tag{5.18}$$

Alors, en utilisant la formule

$$E[X^2] = \mathrm{VAR}[X] + (E[X])^2 \tag{5.19}$$

on trouve que

$$E[S^2] = \frac{1}{n-1}\left\{ \sum_{k=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2 \right) \right\} = \sigma^2$$

$\diamondsuit$

Remarque. Considérons la variable aléatoire discrète X dont la fonction de probabilité est donnée par le tableau suivant:

x	0	1	2
$p_X(x)$	1/3	1/3	1/3

Soit X_1 et X_2 deux observations *différentes* de X; cela correspond à la situation où l'on prélève un échantillon aléatoire de taille 2 *sans* remise d'une population qui compte trois individus. Il y a trois échantillons possibles (si l'on ne tient pas compte de l'ordre): $(0,1)$, $(0,2)$ et $(1,2)$. De plus, puisque $p_X(x)$ est une constante, chaque échantillon a une chance sur trois d'être choisi. Il s'ensuit

que la fonction de probabilité de la moyenne de l'échantillon aléatoire, $\bar{X} := (X_1 + X_2)/2$, est

$\bar{x}$	1/2	1	3/2
$p_{\bar{X}}(\bar{x})$	1/3	1/3	1/3

Maintenant, on trouve facilement que $E[X] \equiv \mu_X = 1$ et $\mathrm{VAR}[X] \equiv \sigma_X^2 = 2/3$. Dans le cas de la variable aléatoire $\bar{X}$, qui n'est *pas* une convolution de X_1 et X_2, on calcule

$$E[\bar{X}] \equiv \mu_{\bar{X}} = \frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + \frac{3}{2} \cdot \frac{1}{3} = 1$$

et

$$\mathrm{VAR}[\bar{X}] \equiv \sigma_{\bar{X}}^2 = \left(\frac{1}{2}\right)^2 \cdot \frac{1}{3} + 1^2 \cdot \frac{1}{3} + \left(\frac{3}{2}\right)^2 \cdot \frac{1}{3} - 1^2 = \frac{14}{4} \cdot \frac{1}{3} - 1 = \frac{1}{6}$$

On a donc: $\mu_{\bar{X}} = \mu_X$ et $\sigma_{\bar{X}}^2 < \sigma_X^2$. En fait, on peut montrer (par induction sur n) que, si l'on prend n individus au hasard et *sans* remise dans une population *finie* comptant N individus (de sorte que $p_X(x) \equiv 1/N$, où X est le numéro de l'individu choisi), alors

$$\mu_{\bar{X}} = \mu_X \quad \text{et} \quad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} \left(\frac{N-n}{N-1}\right)$$

Si on laisse N tendre vers l'infini dans la formule qui donne $\sigma_{\bar{X}}^2$, on trouve que

$$\sigma_{\bar{X}}^2 \longrightarrow \frac{\sigma_X^2}{n}$$

En pratique, si n est inférieur à $0{,}05N$, alors on peut écrire que $\sigma_{\bar{X}}^2 \simeq \sigma_X^2/n$.

Finalement, si l'on prend les observations *avec* remise, ou si la taille N de la population est *infinie*, alors les observations $X_1, \ldots, X_n$ sont *indépendantes* et possèdent la même distribution que la population X. Il s'ensuit que $E[\bar{X}] = E[X]$ et $\mathrm{VAR}[\bar{X}] = \mathrm{VAR}[X]/n$, comme dans l'exemple ci-dessus.

Définition 5.2.5. *On définit l'***erreur quadratique moyenne** *de l'estimateur T d'un paramètre inconnu θ d'une population par*

$$\text{E.Q.M.}[T] = E[(T-\theta)^2] \tag{5.20}$$

Remarques.
(i) En développant le carré dans la définition ci-dessus, on trouve que

$$\text{E.Q.M.}[T] = \text{VAR}[T] + (\text{Biais}[T])^2 \tag{5.21}$$

Donc, si T est un estimateur non biaisé de θ, alors $\text{E.Q.M.}[T] = \text{VAR}[T]$.
(ii) Soit T_1 et T_2 deux estimateurs de θ. On dit que T_1 est *relativement plus efficace* (ou simplement *meilleur*) que T_2 si

$$\text{E.Q.M.}[T_1] < \text{E.Q.M.}[T_2] \tag{5.22}$$

Exemple 5.2.4. Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une population normale X de moyenne μ et de variance σ^2, où les constantes μ et σ^2 sont inconnues. En utilisant la formule (5.17), on peut écrire que

$$\text{E.Q.M.}[\bar{X}] = \text{VAR}[\bar{X}] = \sigma^2/n \tag{5.23}$$

De plus, on peut montrer que

$$\text{E.Q.M.}[S^2] = \frac{2}{n-1}\sigma^4 \tag{5.24}$$

En effet, on verra à la section 3 que, à une constante multiplicative près, la variable aléatoire S^2 présente une distribution du khi-deux:

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

ce qui est un résultat classique en statistique mathématique. Étant donné que la distribution χ^2_{n-1} est identique à la distribution $\text{G}\big(\alpha = \frac{n-1}{2}, \lambda = \frac{1}{2}\big)$, on peut écrire ($S^2$ étant non biaisé pour σ^2) que

$$\begin{aligned}\text{E.Q.M.}[S^2] = \text{VAR}[S^2] &= \frac{\sigma^4}{(n-1)^2}\text{VAR}\left[(n-1)\frac{S^2}{\sigma^2}\right] \\ &= \frac{\sigma^4}{(n-1)^2}\frac{(n-1)/2}{(1/2)^2} = \frac{2}{n-1}\sigma^4\end{aligned}$$

En fait, on trouve que l'estimateur (biaisé) de σ^2 obtenu en divisant par n plutôt que par $n-1$ dans la formule (5.12) est relativement plus efficace que S^2. En effet, si l'on note $\hat{\sigma}^2$ cet estimateur, alors on peut écrire que

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2$$

de sorte que

$$E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \implies \text{Biais}[\hat{\sigma}^2] = -\frac{\sigma^2}{n}$$

et

$$\text{VAR}[\hat{\sigma}^2] = \frac{(n-1)^2}{n^2}\text{VAR}[S^2] = \frac{2(n-1)}{n^2}\sigma^4$$

Ainsi, on calcule

$$\text{E.Q.M.}[\hat{\sigma}^2] = \frac{2(n-1)}{n^2}\sigma^4 + \left(-\frac{\sigma^2}{n}\right)^2 = \frac{2n-1}{n^2}\sigma^4$$

Or, on a:

$$\frac{2n-1}{n^2} < \frac{2}{n-1} \iff 2n^2 - 3n + 1 < 2n^2 \iff 3n > 1$$

ce qui est vrai pour tout $n \in \{2, 3, \ldots\}$. ♢

Remarque. Le critère de l'erreur quadratique moyenne pour décider lequel est le *meilleur*, entre deux ou plusieurs estimateurs d'un paramètre inconnu, est excellent (et très répandu). Cependant, ce critère n'est pas infaillible. Par exemple, définissons l'estimateur $\hat{\sigma}^2(c)$ de la variance σ^2 d'une distribution normale par

$$\hat{\sigma}^2(c) = c(n-1)S^2$$

où c est une constante positive. On trouve que la constante c qui minimise l'erreur quadratique moyenne de $\hat{\sigma}^2(c)$ n'est pas $1/n$, mais plutôt $1/(n+1)$. En effet, on peut vérifier facilement que

$$E[\hat{\sigma}^2(c)] = c(n-1)\sigma^2 \implies \text{Biais}[\hat{\sigma}^2(c)] = [c(n-1)-1]\sigma^2$$

et

$$\text{VAR}[\hat{\sigma}^2(c)] = 2c^2(n-1)\sigma^4 \implies \text{E.Q.M.}[\hat{\sigma}^2(c)] = 2c^2(n-1)\sigma^4 + [c(n-1)-1]^2\sigma^4$$

Enfin, on a:

$$\frac{d}{dc}\text{E.Q.M.}[\hat{\sigma}^2(c)] = 0 \iff c = n+1$$

(et on peut montrer que $c = n + 1$ correspond bien à un *minimum*). Or, il n'y a personne qui utilise

$$\hat{\sigma}_1^2 := \sum_{k=1}^{n} \frac{(X_k - \bar{X})^2}{n+1}$$

comme estimateur de σ^2.

Borne de Cramér-Rao. Soit X une variable aléatoire dont la fonction de densité $f_X(x;\theta)$ ou la fonction de probabilité $p_X(x;\theta)$ dépend d'un paramètre inconnu θ.

Remarque. Nous utilisons la notation $f_X(x;\theta)$ plutôt que $f_X(x,\theta)$ pour indiquer que la fonction f_X est une fonction d'une seule variable, soit x; la quantité θ est une constante qui apparaît dans la fonction f_X. Certains auteurs écrivent $f_X(x \mid \theta)$. De plus, pour simplifier l'écriture, nous allons parfois utiliser f_X indifféremment pour désigner une fonction de densité ou une fonction de probabilité.

On peut montrer que

$$\mathrm{VAR}[T] \geq \frac{1}{nE\left[\frac{d}{d\theta} \ln f_X(x;\theta)\right]^2} \tag{5.25}$$

pour tout estimateur *non biaisé* T de θ. S'il existe un estimateur sans biais de θ qui atteint la borne de Cramér-Rao, on l'appelle **estimateur sans biais à variance minimale**.

Remarques.
(i) On peut aussi montrer que, sous certaines *conditions de régularité*, la borne s'écrit comme suit:

$$\mathrm{VAR}[T] \geq \frac{1}{-nE\left[\frac{d^2}{d\theta^2} \ln f_X(x;\theta)\right]} \tag{5.26}$$

Cette expression est souvent beaucoup plus facile à calculer.

(ii) Pour que le calcul de la borne de Cramér-Rao soit valide, il faut, en particulier, que l'ensemble des valeurs possibles de la variable aléatoire X ne dépende pas du paramètre θ.

(iii) Il serait peut-être plus rigoureux d'écrire $\partial/\partial\theta$ que $d/d\theta$.

Exemple 5.2.5. Soit $X_1, X_2, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X qui présente une distribution de Poisson de paramètre λ. Alors $\bar{X}$ est un estimateur non biaisé de λ, car $E[\bar{X}] = E[X] = \lambda$. Ensuite, on a:

$$\frac{d}{d\lambda} \ln p_X(X;\lambda) = \frac{d}{d\lambda} \ln \left(\frac{e^{-\lambda}\lambda^X}{X!} \right) = \frac{d}{d\lambda}(-\lambda + X \ln \lambda - \ln X!) = -1 + \frac{X}{\lambda}$$

de sorte que

$$E\left[\frac{d}{d\lambda} \ln p_X(x;\lambda) \right]^2 = E\left[\frac{X}{\lambda} - 1 \right]^2 = \frac{1}{\lambda^2} E[(X-\lambda)^2] = \frac{1}{\lambda^2} \mathrm{VAR}[X] = \frac{1}{\lambda}$$

Par conséquent, on peut écrire que

$$\mathrm{VAR}[T] \geq \frac{1}{n\left(\frac{1}{\lambda}\right)} = \frac{\lambda}{n}$$

Or, on a:

$$\mathrm{VAR}[\bar{X}] = \frac{1}{n} \mathrm{VAR}[X] = \frac{\lambda}{n}$$

Donc, on peut conclure que $\bar{X}$ est l'estimateur sans biais à variance minimale de λ.

Notons qu'ici on a:

$$E\left[\frac{d^2}{d\lambda^2} \ln p_X(x;\lambda) \right] = E\left[-\frac{X}{\lambda^2} \right] = -\frac{1}{\lambda}$$

ce qui mène effectivement à la même borne λ/n. ♢

Définition 5.2.6. *Soit T un estimateur de θ basé sur un échantillon aléatoire de taille n. On dit que T est un estimateur* **convergent** *de θ si T converge en probabilité vers θ; c'est-à-dire si*

$$\lim_{n\to\infty} P[|T-\theta| \leq \epsilon] = 1 \quad \forall \epsilon > 0 \tag{5.27}$$

Exemple 5.2.6. La moyenne de l'échantillon aléatoire, $\bar{X}$, est un estimateur convergent de μ, car on peut écrire, par la loi (faible) des grands nombres, que

$$\lim_{n\to\infty} P\left[|\bar{X}-\mu| \leq \epsilon\right] = \lim_{n\to\infty} \left\{1 - P\left[|\bar{X}-\mu| > \epsilon\right]\right\} = 1 - 0 = 1$$

pour toute constante $\epsilon > 0$.

Notons que l'on peut aussi utiliser l'inégalité de Bienaymé-Tchebychev pour montrer que $\bar{X}$ est un estimateur convergent de μ. ♢

5.2.2 La méthode du maximum de vraisemblance

Définition 5.2.7. *Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une population X dont la fonction de densité (ou la fonction de probabilité) $f_X(x;\theta)$ dépend d'un paramètre inconnu θ. La* **fonction de vraisemblance** *de l'échantillon est*

$$L(\theta) = \prod_{k=1}^{n} f_X(X_k;\theta) \tag{5.28}$$

Définition 5.2.8. *L'***estimateur à vraisemblance maximale**, *θ_{VM}, du paramètre θ est la valeur de θ qui maximise la fonction de vraisemblance.*

Remarques.
(i) On peut avoir: $\theta = (\theta_1, \ldots, \theta_m)$. Dans ce cas, on peut dériver $L(\theta)$ par rapport aux m paramètres, poser que $\frac{\partial}{\partial\theta_m} L(\theta) = 0\ \forall m$ et essayer de résoudre le système d'équations obtenu.

(ii) Soit $\hat{\theta} = \theta_{VM}$, l'estimateur à vraisemblance maximale de θ. On peut montrer que si n est assez grand, alors $\hat{\theta}$ présente approximativement une distribution normale de moyenne θ et de variance

$$\sigma^2_{\hat{\theta}} = \frac{1}{nE\left[\frac{d}{d\theta}\ln f_X(x;\theta)\right]^2} \tag{5.29}$$

C'est-à-dire que le biais de $\hat{\theta}$ décroît vers 0 (si $\hat{\theta}$ est biaisé) et sa variance converge vers la borne de Cramér-Rao.

(iii) On peut aussi montrer que θ_{VM} est un estimateur *convergent* de θ.

(iv) Finalement, si $g(\theta)$ est une fonction de θ qui possède une fonction inverse, alors $g(\theta_{VM})$ est l'estimateur à vraisemblance maximale de $g(\theta)$.

Exemple 5.2.7. (a) Soit $X_1, \ldots, X_n$ un échantillon aléatoire de $X \sim \text{Exp}(\lambda)$. Alors on a:

$$L(\lambda) = \prod_{k=1}^{n} \lambda e^{-\lambda X_k} = \lambda^n e^{-\lambda n \bar{X}} \quad \text{si } X_k \geq 0\ \ \forall k$$

Puisque λ maximise $L(\lambda)$ si et seulement si λ maximise $\ln L(\lambda)$, on peut considérer

$$\ln L(\lambda) = n \ln \lambda - \lambda n \bar{X}$$

On a:

$$\frac{d}{d\lambda}\ln L(\lambda) = \frac{n}{\lambda} - n\bar{X} = 0 \iff \lambda = \frac{1}{\bar{X}}$$

Étant donné que

$$\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{n}{\lambda^2} < 0$$

on peut écrire que

$$\lambda_{VM} = 1/\bar{X}$$

(b) Si $X \sim \text{Poi}(\lambda)$, alors on a:

$$L(\lambda) = \prod_{k=1}^{n} \frac{e^{-\lambda}\lambda^{X_k}}{X_k!} = \frac{e^{-n\lambda}\lambda^{n\bar{X}}}{\prod_{k=1}^{n} X_k!} \quad \text{si } X_k \in \{0, 1, \ldots\} \ \ \forall k$$

$$\implies \ln L(\lambda) = -n\lambda + n\bar{X}\ln\lambda - \sum_{k=1}^{n}\ln X_k!$$

$$\implies \frac{d}{d\lambda}\ln L(\lambda) = -n + \frac{n\bar{X}}{\lambda} = 0 \iff \lambda = \bar{X}$$

Donc, on conclut que

$$\lambda_{VM} = \bar{X}$$

car

$$\frac{d^2}{d\lambda^2}\ln L(\lambda) = -\frac{n\bar{X}}{\lambda^2} < 0$$

(au moins si n est assez grand, puisque la moyenne $\bar{X}$ est strictement positive dès qu'une observation est positive). ♢

Remarques.
(i) **Méthode des moments**. Cette méthode consiste à poser l'égalité entre les moments (théoriques) de la variable aléatoire X et les *moments de l'échantillon*. Ainsi, dans le cas où $X \sim \text{Exp}(\lambda)$, on pose que

$$E[X] = \bar{X} \iff \frac{1}{\lambda} = \bar{X}$$

Il s'ensuit que l'estimateur du paramètre λ par la méthode des moments, λ_M, est aussi donné par $1/\bar{X}$. De même, lorsque $X \sim \text{Poi}(\lambda)$, on trouve immédiatement que $E[X] = \lambda \Rightarrow \lambda_M = \bar{X}$ $(= \lambda_{VM})$.

En général, on pose:

$$E[X^m] = \sum_{k=1}^{n} \frac{X_k^m}{n} \quad \text{pour } m = 1, 2, \ldots \tag{5.30}$$

On se sert de j équations (utiles) pour estimer les j paramètres inconnus de la fonction $f_X(x;\theta)$. Par exemple, si $X \sim \mathrm{U}[-\theta, \theta]$, alors on utilise l'équation obtenue avec $m = 2$ (car l'équation avec $m = 1$ ne permet pas d'estimer θ):

$$E[X^2] = \sum_{k=1}^{n} \frac{X_k^2}{n}$$

C'est-à-dire

$$\frac{(2\theta)^2}{12} = \sum_{k=1}^{n} \frac{X_k^2}{n} \implies \theta_M = \left[3 \sum_{k=1}^{n} \frac{X_k^2}{n}\right]^{1/2} \tag{5.31}$$

En fait, certains auteurs utilisent plutôt l'équation

$$\mathrm{VAR}[X] = S^2 \iff \mathrm{VAR}[X] = \sum_{k=1}^{n} \frac{(X_k - \bar{X})^2}{n-1}$$

pour estimer le paramètre θ, ce qui donne un estimateur différent:

$$\hat{\theta} = \left[3 \sum_{k=1}^{n} \frac{(X_k - \bar{X})^2}{n-1}\right]^{1/2}$$

Cependant, étant donné que $E[X] = 0$, si n est assez grand, alors $\bar{X}$ devrait être environ égal à zéro, de sorte que les deux estimateurs donneront des valeurs presque égales.

(ii) Soit $X \sim \mathrm{U}[0, \theta]$. Alors

$$L(\theta) = \frac{1}{\theta^n} \quad \text{si } X_k \in [0, \theta] \ \ \forall k \tag{5.32}$$

Dans ce cas, on ne peut pas obtenir la valeur de θ qui maximise $L(\theta)$ en dérivant, car l'ensemble des valeurs possibles de X dépend du paramètre inconnu θ. Cependant, puisque $L(\theta)$ est une fonction strictement décroissante, et puisque l'on doit avoir: $\theta \geq X_k \ \forall k$, on peut conclure que

$$\theta_{VM} = \max\{X_1, \ldots, X_n\}$$

ce qui est la plus petite valeur que θ peut prendre si l'on a obtenu les observations $X_1, \ldots, X_n$.

Pour obtenir l'estimateur du paramètre θ par la méthode des moments, on pose:

$$E[X] = \bar{X} \iff \frac{\theta}{2} = \bar{X} \implies \theta_M = 2\bar{X}$$

Notons que cet estimateur de θ *peut* donner des résultats qui n'ont pas de sens. Par exemple, si on a recueilli les données suivantes:

$$x_1 = 1, \quad x_2 = 2 \quad \text{et} \quad x_3 = 9$$

alors l'estimation ponctuelle de θ, selon la méthode de vraisemblance maximale, est $\hat{\theta} = 9$, tandis que $\hat{\theta} = 2 \times 4 = 8$ selon la méthode des moments. Or, si l'on a obtenu $x_3 = 9$, alors le paramètre θ doit être au moins égal à 9. Par conséquent, l'estimation ponctuelle $\hat{\theta} = 8$ est clairement mauvaise. Par contre, si l'on a plutôt observé

$$x_1 = 1, \quad x_2 = 2 \quad \text{et} \quad x_3 = 3$$

alors on a: $\hat{\theta}_{VM} = 3$ et $\hat{\theta}_M = 4$. Même si la valeur $\hat{\theta}_{VM} = 3$ est théoriquement possible, il est logique de croire que θ doit être supérieur à $x_{(3)} = 3$. De plus, l'estimateur θ_M possède une propriété importante: sa distribution tend vers une distribution normale. Donc, la méthode des moments vaut la peine d'être considérée pour estimer les paramètres inconnus.

(iii) **Méthode des moindres carrés.** Pour conclure cette section, nous allons décrire brièvement la méthode d'estimation des paramètres appelée *méthode des moindres carrés*, que nous utiliserons au chapitre 7, portant sur la *régression*. (En fait, dans le problème considéré au chapitre 7, la méthode des moindres carrés sera équivalente à celle du maximum de vraisemblance.)

Soit $Y_1, Y_2, \ldots, Y_n$ des variables aléatoires *indépendantes*. Supposons que

$$E[Y_i] = \sum_{j=1}^{k} \theta_j x_{ij} \quad \text{pour } i = 1, 2, \ldots, n \tag{5.33}$$

où les θ_j sont des paramètres inconnus et les x_{ij} sont des quantités connues. Les *estimateurs des moindres carrés* $\hat{\theta}_1, \ldots, \hat{\theta}_k$ des paramètres $\theta_1, \ldots, \theta_k$ sont les valeurs qui minimisent la fonction

$$SC(\theta_1, \ldots, \theta_k) := \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} \theta_j x_{ij} \right)^2 \quad (\text{où } n > k) \tag{5.34}$$

Pour obtenir $\hat{\theta}_1, \ldots, \hat{\theta}_k$, on peut résoudre le système

$$\sum_{i=1}^{n} x_{im} \left(Y_i - \sum_{j=1}^{k} \hat{\theta}_j x_{ij} \right) = 0 \quad \text{pour } m = 1, \ldots, k \tag{5.35}$$

Exemple 5.2.8. Supposons que

$$Y_i = \theta_1 x_i + \epsilon_i$$

pour $i = 1, \ldots, n$, où ϵ_i est une variable aléatoire telle que $E[\epsilon_i] = 0 \ \forall i$. On a:

$$\frac{d}{d\theta_1} SC(\theta_1) = \frac{d}{d\theta_1} \sum_{i=1}^{n} (Y_i - \theta_1 x_i)^2 = \sum_{i=1}^{n} 2(Y_i - \theta_1 x_i)(-x_i)$$

On trouve que

$$SC'(\theta_1) = 0 \iff \sum_{i=1}^{n} Y_i x_i = \theta_1 \sum_{i=1}^{n} x_i^2$$

de sorte que

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} Y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

♢

5.3 Distributions d'échantillonnage

Comme nous l'avons vu à la section précédente, la distribution que présente une statistique est appelée *loi* ou *distribution d'échantillonnage*. Nous avons aussi vu, dans l'exemple 5.2.1, que la distribution normale est une distribution d'échantillonnage, car la moyenne d'un échantillon aléatoire de taille n d'une variable aléatoire X qui présente une distribution normale, de moyenne μ et de variance σ^2, présente une distribution $\mathrm{N}(\mu, \sigma^2/n)$. De plus, ce résultat est approximativement vrai, par le théorème central limite, même si X ne présente pas une distribution normale, pourvu que n soit assez grand.

Dans cette section, nous allons définir les autres distributions d'échantillonnage les plus importantes.

Définition 5.3.1. *Soit $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes, où $X_i \sim N(\mu_i, \sigma_i^2)$ pour tout i. On dit que la variable aléatoire*

$$X := \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2 \tag{5.36}$$

présente une **distribution du khi-deux** *ou* **khi-carré** *à n* **degrés de liberté**. *On écrit: $X \sim \chi_n^2$.*

Remarques.
(i) Le nombre de degrés de liberté est le nombre de variables aléatoires *indépendantes* que l'on additionne.

(ii) Notons que le carré d'une variable aléatoire $\mathrm{N}(0,1)$ présente une distribution χ_1^2.

Propriétés.

(i) On peut montrer que la fonction de densité de X est donnée par

$$f_X(x) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2} \quad \text{pour } x > 0 \tag{5.37}$$

(ii) Nous avons mentionné, à la section 3.4, que la distribution du khi-deux est un cas particulier de la distribution gamma. Plus précisément, $X \sim \chi_n^2 \Leftrightarrow X \sim G(\alpha = \frac{n}{2}, \lambda = \frac{1}{2})$. Il s'ensuit que

$$E[X] = \frac{\alpha}{\lambda} = n \quad \text{et} \quad \mathrm{VAR}[X] = \frac{\alpha}{\lambda^2} = 2n \tag{5.38}$$

Notons que les formules pour la moyenne et la variance de X se déduisent aussi facilement de la définition de X en fonction des variables aléatoires X_i (voir la formule (5.36)).

(iii) En utilisant le théorème central limite, on peut affirmer que la fonction de densité d'une distribution du khi-deux tend vers celle d'une distribution *normale* lorsque n tend vers l'infini.

(iv) Soit $X_1, X_2, \ldots, X_n$ n variables aléatoires *indépendantes*, où $X_i \sim \chi_{\nu_i}^2$ pour $i = 1, 2, \ldots, n$. Alors il est facile de démontrer que la variable aléatoire $Y := \sum_{i=1}^{n} X_i \sim \chi_\nu^2$, où $\nu := \sum_{i=1}^{n} \nu_i$.

Définition 5.3.2. *Soit* $Z \sim \mathrm{N}(0,1)$ *et* $Y \sim \chi^2_n$ *deux variables aléatoires indépendantes. On dit que la variable aléatoire*

$$X := \frac{Z}{\sqrt{Y/n}} \tag{5.39}$$

présente une **distribution** t **de Student** *à* n **degrés de liberté**. *On écrit:* $X \sim t_n$.

Remarque. Student est le pseudonyme de W. S. Gosset (voir la section 1.2).

Propriétés.

(i) On peut montrer que

$$f_X(x) = \frac{1}{\sqrt{\pi n}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \quad \text{pour } x \in \mathbb{R} \tag{5.40}$$

(ii) La fonction f_X est *symétrique* par rapport à l'origine et a la même allure que la densité de $Z \sim \mathrm{N}(0,1)$ (voir la figure 5.3). En fait, si $n \to \infty$, alors $f_X(x)$ tend vers la densité de Z.

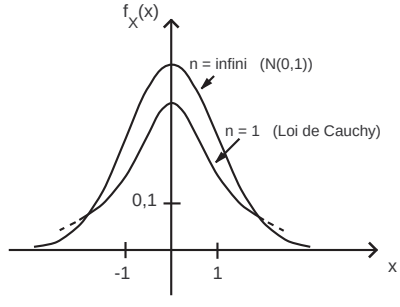

Fig. 5.3. Exemples de distributions de Student

(iii) La moyenne et la variance de $X \sim t_n$ sont

$$E[X] = 0 \quad (\text{si } n > 1) \quad \text{et} \quad \mathrm{VAR}[X] = \frac{n}{n-2} \quad (\text{si } n > 2) \tag{5.41}$$

Pour $n = 1$, la moyenne et la variance de X n'existent pas, tandis que pour $n = 2$, on a: $E[X] = 0$, mais $\mathrm{VAR}[X] = E[X^2] = \infty$.

(iv) Si $n = 1$, on a:

$$f_X(x) = \frac{1}{\pi}(1 + x^2)^{-1} \quad \text{pour } x \in \mathbb{R} \tag{5.42}$$

Cette variable aléatoire porte aussi le nom de **distribution de Cauchy**.

Définition 5.3.3. *Soit $V \sim \chi^2_m$ et $W \sim \chi^2_n$ deux variables aléatoires indépendantes. On dit que la variable aléatoire (positive)*

$$X := \frac{V/m}{W/n} \tag{5.43}$$

présente une **distribution F de Fisher** *à m* **degrés de liberté** *au numérateur et n* **degrés de liberté** *au dénominateur. On écrit: $X \sim F_{m,n}$.*

Propriétés.

(i) On peut montrer que

$$f_X(x) = \frac{\Gamma(\frac{m+n}{2})\left(\frac{m}{n}\right)^{\frac{m}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{x^{\frac{m}{2}-1}}{\left[\left(\frac{m}{n}\right)x + 1\right]^{\frac{m+n}{2}}} \quad \text{pour } x > 0 \tag{5.44}$$

(voir la figure 5.4).

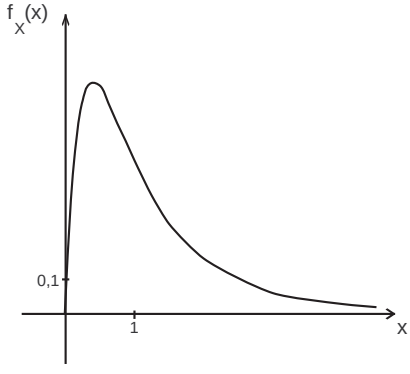

Fig. 5.4. Fonction de densité de la distribution de Fisher avec $m = 4$ et $n = 8$

(ii) On déduit directement de la définition que $Y := 1/X \sim F_{n,m}$. On trouve aussi que si T est une variable aléatoire qui présente une distribution t_n, alors $T^2 \sim F_{1,n}$ (car si $Z \sim \mathrm{N}(0, 1)$, alors $Z^2 \sim \chi^2_1$).

(iii) Si $X \sim F_{m,n}$, alors

$$E[X] = \frac{n}{n-2} \quad (\text{si } n > 2) \quad \text{et} \quad \text{VAR}[X] = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \quad (\text{si } n > 4) \tag{5.45}$$

Nous allons maintenant donner plusieurs exemples (qui seront utiles par la suite) de statistiques qui présentent l'une ou l'autre des distributions d'échantillonnage définies ci-dessus.

Si $X_1, \ldots, X_n$ est un échantillon aléatoire de la variable aléatoire X qui présente une distribution $\text{N}(\mu, \sigma^2)$, alors le résultat

$$\sum_{i=1}^{n} X_i \sim \text{N}(n\mu, n\sigma^2) \implies \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{N}(0,1) \tag{5.46}$$

est *exact* pour n'importe quelle valeur de n. De plus, on sait que

$$\frac{X_i - \mu}{\sigma} \sim \text{N}(0,1) \quad \text{pour } i = 1, \ldots, n \tag{5.47}$$

Ainsi, par indépendance, on peut écrire que

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n \tag{5.48}$$

Si μ est inconnu, alors il faudra le remplacer par son estimateur $\bar{X}$ dans les formules donnant des *intervalles de confiance* et dans les *tests d'hypothèses*. On peut montrer que cela a pour effet de faire perdre un degré de liberté. C'est-à-dire que

$$\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1} \tag{5.49}$$

Finalement, il s'ensuit que

$$(n-1)\frac{S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1} \tag{5.50}$$

Si l'écart-type σ de X est inconnu, alors il faudra le remplacer par son estimateur S. Nous utiliserons la quantité

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

On peut écrire que

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\left\{\frac{(n-1)(S/\sigma)^2}{n-1}\right\}^{1/2}} \tag{5.51}$$

Or, on peut montrer que $\bar{X}$ et S sont des variables aléatoires indépendantes. Puisque le numérateur dans le membre droit de l'équation (5.51) présente une distribution $\mathrm{N}(0,1)$, alors, par définition (et (5.50)),

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1} \tag{5.52}$$

De plus, même si X ne présente pas une distribution normale, si la taille n de l'échantillon aléatoire est assez grande, alors on peut écrire que

$$\bar{X} \sim \mathrm{N}(\mu, S^2/n) \quad \text{lorsque } n \to \infty \tag{5.53}$$

(en supposant que l'écart-type σ de X est inconnu).

À partir du résultat (5.50), on déduit que, si S_1^2 et S_2^2 sont les variances de deux échantillons aléatoires provenant de populations normales *indépendantes*, alors

$$(n_i-1)\frac{S_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1} \quad \text{pour } i = 1,2$$

où n_i est la taille de l'échantillon aléatoire de la population i, et

$$\frac{(n_1-1)\frac{S_1^2}{\sigma_1^2}/(n_1-1)}{(n_2-1)\frac{S_2^2}{\sigma_2^2}/(n_2-1)} \sim F_{n_1-1,n_2-1} \tag{5.54}$$

C'est-à-dire que l'on a:

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1} \tag{5.55}$$

Donc, si l'on suppose que les deux variances *théoriques* σ_1^2 et σ_2^2 sont égales, alors on obtient que $S_1^2/S_2^2 \sim F_{n_1-1,n_2-1}$.

On utilisera les résultats (5.50) et (5.55) pour obtenir des intervalles de confiance et pour effectuer des tests d'hypothèses au sujet des paramètres σ^2, σ_1^2, σ_2^2.

Exemple 5.3.1. Soit X_1, X_2, X_3 un échantillon aléatoire de $X \sim \mathrm{N}(0,1)$ et Y_1, Y_2, Y_3 un échantillon aléatoire de $Y \sim \mathrm{N}(12,4)$. On suppose que X et Y sont des variables aléatoires indépendantes. On définit

$$U = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2$$

et

$$V = \frac{(Y_1 - 12)^2}{4} + \frac{(Y_2 - 12)^2}{4} + \frac{(Y_3 - 12)^2}{4}$$

(a) Trouver un nombre a tel que $P[U \leq a] \simeq 0{,}95$.
(b) Trouver un nombre b tel que $P\left[\frac{U}{V} \leq b\right] \simeq 0{,}025$.

Solution. (a) On a:

$$S_X^2 := \sum_{i=1}^{3} \frac{(X_i - \bar{X})^2}{3-1} = \frac{\sigma_X^2}{2} \sum_{i=1}^{3} \frac{(X_i - \bar{X})^2}{\sigma_X^2}$$

Alors, étant donné que $\sigma_X^2 = 1$, on peut écrire que

$$U = \frac{2S_X^2}{\sigma_X^2} \sim \chi^2_{3-1}$$

Maintenant, on trouve dans le tableau 5.3, à la page 240, que $P[X \geq 5{,}99] \simeq 0{,}05$ si $X \sim \chi_2^2$. Ce tableau donne les nombres $\chi^2_{\alpha,n}$ pour lesquels la probabilité qu'une variable aléatoire X, qui présente une distribution χ_n^2, soit supérieure ou égale à $\chi^2_{\alpha,n}$ est égale à α (voir la section 5.4). Il s'ensuit que $a = 5{,}99$, puisque

$$P[U \leq 5{,}99] = 1 - P[\chi_2^2 > 5{,}99] \simeq 1 - 0{,}05 = 0{,}95$$

(b) On a:

$$\frac{(Y_i - 12)^2}{4} \sim \chi_1^2$$

et, par indépendance,

$$\sum_{i=1}^{3} \frac{(Y_i - 12)^2}{4} \sim \chi_3^2$$

Alors

$$P\left[\frac{U}{V} \leq b\right] = P\left[\frac{U/2}{V/3} \leq \frac{3b}{2}\right] \overset{\text{ind.}}{=} P\left[F_{2,3} \leq \frac{3b}{2}\right]$$

De plus, nous avons vu que si $X \sim F_{m,n}$, alors $1/X \sim F_{n,m}$; il s'ensuit que

$$P\left[F_{2,3} \leq \frac{3b}{2}\right] = P\left[F_{3,2} \geq \frac{2}{3b}\right] \simeq 0{,}025 \iff \frac{2}{3b} \simeq 39{,}2$$

d'après le tableau des valeurs de $F_{0,025;n_1,n_2}$ à l'appendice B, à la page 519, pour lesquelles

$$P[F_{n_1,n_2} \geq F_{0,025;n_1,n_2}] = 0{,}025$$

Donc, on trouve que $b \simeq 0{,}0255$.

Remarque. Si l'on définit plutôt

$$V = \sum_{i=1}^{3} \frac{(Y_i - \bar{Y})^2}{4}$$

alors on a: $V \sim \chi^2_2$ et $U/V \sim F_{2,2}$. Il s'ensuit que

$$\frac{1}{b} \simeq 39{,}0 \iff b \simeq 0{,}0256$$

car $P[F_{2,2} \geq 39{,}0] \simeq 0{,}025$, selon le tableau 5.4, à la page 243 (ou le tableau des $F_{0,025;n_1,n_2}$, à l'appendice B). ♢

5.4 Estimation par intervalles de confiance

Définition 5.4.1. *Un intervalle* $[T_I, T_S]$ *est appelé* **intervalle de confiance (bilatéral)** *à* $100(1-\alpha)$ % *pour le paramètre* θ *si*

$$P[T_I(X_1, \ldots, X_n) \leq \theta \leq T_S(X_1, \ldots, X_n)] = 1 - \alpha \tag{5.56}$$

Remarques.

(i) On appelle T_I et T_S **limites inférieure** et **supérieure de confiance**, respectivement; de plus, $1-\alpha$ est appelé **coefficient de confiance**.

(ii) θ est un paramètre, tandis que T_I et T_S sont des variables aléatoires.

(iii) Des intervalles de confiance **unilatéraux** avec **borne inférieure** et **borne supérieure** sont donnés respectivement par $[T_I, \infty)$ et $(-\infty, T_S]$, où maintenant

$$P[T_I \leq \theta] = P[\theta \leq T_S] = 1 - \alpha \tag{5.57}$$

Notons que, si l'on désire construire un intervalle de confiance avec borne supérieure pour une variance σ^2, par exemple, alors l'intervalle sera plutôt de la forme $[0, T_S]$.

(iv) Les intervalles de confiance les plus courts sont souvent les intervalles de confiance bilatéraux *symétriques*; c'est-à-dire ceux pour lesquels on a:

$$P[\theta < T_I] = P[\theta > T_S] = \alpha/2$$

5.4.1 Intervalle de confiance pour μ; σ connu

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une population normale X, dont la moyenne μ est inconnue, mais la variance σ^2 est connue. La technique pour obtenir un intervalle de confiance pour un paramètre (inconnu) consiste d'abord à trouver une quantité qui présente une des distributions d'échantillonnage que nous avons vues. La seule inconnue dans cette quantité doit être le paramètre qui nous intéresse. Ici, on peut écrire que $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$, ce qui implique que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1) \tag{5.58}$$

Il s'ensuit que

$$P\left[- z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} \right] = 1 - \alpha \tag{5.59}$$

où $z_{\alpha/2}$ est défini par (voir la figure 5.5)

$$\Phi(z_{\alpha/2}) = 1 - \alpha/2 \text{ ou } Q(z_{\alpha/2}) = \alpha/2 \tag{5.60}$$

C'est-à-dire que $z_{0,05}$, par exemple, est le 95$^{\mathrm{e}}$ centile de la distribution normale centrée réduite.

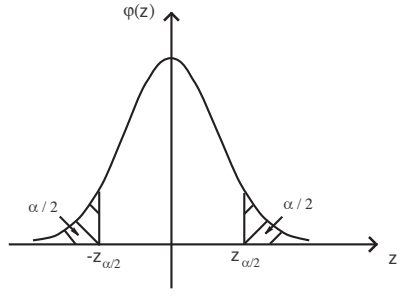

Fig. 5.5. Définition de la quantité $z_{\alpha/2}$ pour une variable aléatoire Z qui présente une distribution normale centrée réduite

On déduit de l'équation (5.59) que

$$P\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha \tag{5.61}$$

Ainsi, l'intervalle

$$\left[\bar{X} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{X} + z_{\alpha/2}(\sigma/\sqrt{n})\right] \tag{5.62}$$

est un intervalle de confiance à $100(1-\alpha)$ % pour μ. Lorsqu'un échantillon aléatoire *particulier* de la population X a été recueilli, on obtient une réalisation de l'intervalle de confiance en calculant

$$\left[\bar{x} - z_{\alpha/2}(\sigma/\sqrt{n}), \bar{x} + z_{\alpha/2}(\sigma/\sqrt{n})\right] \tag{5.63}$$

Remarques.
(i) Des intervalles de confiance unilatéraux à $100(1-\alpha)$ % pour μ sont donnés par

$$\left[\bar{X} - z_{\alpha}(\sigma/\sqrt{n}), \infty\right) \quad \text{et} \quad \left(-\infty, \bar{X} + z_{\alpha}(\sigma/\sqrt{n})\right] \tag{5.64}$$

Dans le cas bilatéral, on écrit souvent l'intervalle de confiance sous forme compacte comme suit:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{5.65}$$

(ii) Les intervalles de confiance donnés par les formules (5.62) et (5.64) sont encore valides (approximativement) même si X ne présente pas une distribution normale, pourvu que n soit suffisamment grand (par le théorème central limite).
(iii) Les valeurs de z_α peuvent être obtenues à l'aide d'une calculatrice ou trouvées dans une table statistique. Les valeurs les plus utiles sont données dans le tableau 5.1. De plus, par symétrie, on a: $z_{1-\alpha} = -z_\alpha$. Finalement, il existe

Tableau 5.1. Valeurs de z_α

α	0,25	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
z_α	0,674	1,282	1,645	1,960	2,326	2,576	3,090	3,291

aussi des formules qui donnent de bonnes approximations de z_α.

Exemple 5.4.1. Soit $x_1, \ldots, x_{10}$ un échantillon aléatoire particulier d'une population $X \sim \mathrm{N}(\mu, \sigma^2 = 0{,}04)$. Supposons que $\bar{x} = 10{,}5$; alors un intervalle de confiance avec coefficient de confiance 95 % pour μ est donné par

$$10,5 \pm 1{,}960(0{,}2/\sqrt{10}) \implies [10{,}38; 10{,}62] \quad \text{(environ)}$$

Remarque. On ne peut pas écrire, comme on pourrait être tenté de le faire, que $P[\mu \in [10{,}38; 10{,}62]] \simeq 0{,}95$, car μ n'est pas une variable aléatoire. Une fois que les données ont été recueillies, on calcule un intervalle *déterministe*; alors le paramètre μ *est* ou n'*est pas* dans cet intervalle. Il n'y a plus de probabilité. $\diamondsuit$

L'erreur maximale (en valeur absolue) que l'on commet, lorsqu'on estime le paramètre inconnu μ par la moyenne d'un échantillon aléatoire de taille n, est

$$E_{max} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{5.66}$$

et ce, avec un coefficient de confiance de $100(1-\alpha)$ %. Par conséquent, si l'on veut que cette erreur maximale soit égale à une constante donnée, alors il faut prendre un échantillon de taille

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E_{max}}\right)^2 \tag{5.67}$$

Puisque cette formule ne donnera pas un entier, en général, il faut prendre le plus petit entier supérieur ou égal à la valeur obtenue avec (5.67). Notons aussi que la longueur de l'intervalle de confiance est égale à $2E_{max}$.

5.4.2 Intervalle de confiance pour μ; σ inconnu

Si la variance σ^2 de la population X est inconnue, mais si n est suffisamment grand, alors on peut utiliser les formules (5.62) et (5.64) tout de même. Il suffit de remplacer σ par son estimateur S. Cependant, si n est petit (< 30), alors les formules précédentes ne sont plus valides. Dans ce cas, il *faut* que X présente une distribution normale. On considère

$$T := \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{5.68}$$

Nous avons vu, à la section précédente, que T présente une distribution t de Student à $n-1$ degrés de liberté. En refaisant le même travail que dans le cas où σ est connu, on trouve que

$$[\bar{X} - t_{\alpha/2,n-1}(S/\sqrt{n}), \bar{X} + t_{\alpha/2,n-1}(S/\sqrt{n})] \tag{5.69}$$

est un intervalle de confiance à $100(1-\alpha)$ % pour μ, où $t_{\alpha/2,n-1}$ est défini par (voir la figure 5.6)

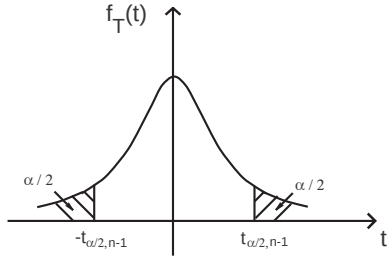

Fig. 5.6. Définition de la quantité $t_{\alpha/2,n-1}$ pour une variable aléatoire T qui présente une distribution t_{n-1}

$$P[T \leq t_{\alpha/2,n-1}] = 1 - \alpha/2 \quad \text{si } T \sim t_{n-1} \tag{5.70}$$

Remarques.
(i) Les formules pour les intervalles de confiance unilatéraux se déduisent facilement de la formule (5.69); on obtient:

$$[\bar{X} - t_{\alpha,n-1}(S/\sqrt{n}), \infty) \quad \text{et} \quad (-\infty, \bar{X} + t_{\alpha,n-1}(S/\sqrt{n})] \tag{5.71}$$

(ii) Les valeurs de $t_{\alpha,n}$ sont obtenues à l'aide d'une calculatrice ou trouvées dans une table. Le tableau 5.2, à la page 236, donne les $t_{\alpha,n}$ pour les principales valeurs de α et pour plusieurs n. De plus, par symétrie, comme dans le cas de la distribution normale centrée réduite, on a: $t_{1-\alpha,n} = -t_{\alpha,n}$.

Tableau 5.2. Valeurs de $t_{\alpha,n}$

n	1	2	3	4	5	6	7	8
$t_{0,005;n}$	63,657	9,925	5,841	4,604	4,032	3,707	3,499	3,355
$t_{0,01;n}$	31,821	6,965	4,541	3,747	3,365	3,143	2,998	2,896
$t_{0,025;n}$	12,706	4,303	3,182	2,776	2,571	2,447	2,365	2,306
$t_{0,05;n}$	6,314	2,920	2,353	2,132	2,015	1,943	1,895	1,860
$t_{0,10;n}$	3,078	1,886	1,638	1,533	1,476	1,440	1,415	1,397

n	9	10	15	20	25	30	40	∞
$t_{0,005;n}$	3,250	3,169	2,947	2,845	2,787	2,750	2,704	2,576
$t_{0,01;n}$	2,821	2,764	2,602	2,528	2,485	2,457	2,423	2,326
$t_{0,025;n}$	2,262	2,228	2,131	2,086	2,060	2,042	2,021	1,960
$t_{0,05;n}$	1,833	1,812	1,753	1,725	1,708	1,697	1,684	1,645
$t_{0,10;n}$	1,383	1,372	1,341	1,325	1,316	1,310	1,303	1,282

Exemple 5.4.2. Si σ est inconnu dans l'exemple 5.4.1 et si l'écart-type de l'*échantillon* est de 0,25, alors l'intervalle de confiance devient

$$10{,}5 \pm 2{,}262(0{,}25/\sqrt{10}) \implies [10{,}32; 10{,}68] \quad \text{(environ)}$$

Remarque. Étant donné que $t_{\alpha,n} > z_\alpha$ pour tout $n \in \{1, 2, \ldots\}$, l'intervalle de confiance obtenu lorsque la variance de la population est inconnue devrait généralement être plus large que l'intervalle de confiance correspondant calculé avec σ connu, ce qui est logique. Cependant, il peut arriver, en pratique, que l'estimateur S de σ sous-estime la vraie valeur de l'écart-type de la population, de sorte que l'intervalle de confiance dans le cas où σ est inconnu est moins large. Par exemple, si l'on avait obtenu une valeur de s égale à 0,15 plutôt que 0,25 ci-dessus, alors l'intervalle de confiance aurait été [10,39; 10,61] (environ), ce qui est un intervalle légèrement moins large que celui calculé dans l'exemple 5.4.1. ♢

Pour conclure cette section, nous allons donner (en omettant les détails) des intervalles de confiance pour d'autres paramètres.

5.4.3 Intervalles de confiance pour $\mu_X - \mu_Y$

Soit $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ des échantillons aléatoires des variables aléatoires *indépendantes* $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, respectivement.

(a) **Cas où μ_X et μ_Y sont inconnus, mais σ_X et σ_Y sont connus**

En utilisant le fait que

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right) \tag{5.72}$$

on trouve que

$$\left[\bar{X} - \bar{Y} - z_{\alpha/2}\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)^{1/2}, \bar{X} - \bar{Y} + z_{\alpha/2}\left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)^{1/2}\right] \tag{5.73}$$

est un intervalle de confiance à $100(1-\alpha)$ % pour $\mu_X - \mu_Y$.

(b) **Cas où tous les paramètres sont inconnus, mais $\sigma_X = \sigma_Y$**

On définit la statistique

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \tag{5.74}$$

où S_X^2 et S_Y^2 sont les variances des échantillons aléatoires. On peut montrer que

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{S_p\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}} \sim t_{m+n-2} \tag{5.75}$$

Il s'ensuit qu'un intervalle de confiance à $100(1-\alpha)$ % pour $\mu_X - \mu_Y$ est donné par

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2,m+n-2} S_p\left(\tfrac{1}{m} + \tfrac{1}{n}\right)^{1/2}, \bar{X} - \bar{Y} + t_{\alpha/2,m+n-2} S_p\left(\tfrac{1}{m} + \tfrac{1}{n}\right)^{1/2}\right] \tag{5.76}$$

(c) **Cas où tous les paramètres sont inconnus, mais m et n sont grands**

Si m et n sont assez grands (≥ 30), alors on peut montrer que

$$\left[\bar{X} - \bar{Y} - t_{\alpha/2,\nu}\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^{1/2}, \bar{X} - \bar{Y} + t_{\alpha/2,\nu}\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^{1/2}\right] \tag{5.77}$$

est un intervalle de confiance *approximatif* à $100(1-\alpha)$ % pour $\mu_X - \mu_Y$, où

$$\nu := \frac{\left(\frac{S_X^2}{m} + \frac{S_Y^2}{n}\right)^2}{\frac{(S_X^2/m)^2}{m-1} + \frac{(S_Y^2/n)^2}{n-1}} \tag{5.78}$$

Notons qu'en général la formule ci-dessus ne donnera pas un entier. On peut arrondir le nombre obtenu ou encore prendre sa partie entière (pour être plus prudent).

5.4.4 Intervalles de confiance pour σ^2

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X qui présente une distribution $\mathrm{N}(\mu, \sigma^2)$.

(a) **Cas où μ est inconnu**

Un résultat important en statistique mathématique, que nous avons vu à la section précédente (voir la page 229), est le suivant: lorsque $X \sim \mathrm{N}(\mu, \sigma^2)$, on trouve que

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{5.79}$$

où S^2 est la variance d'un échantillon aléatoire de taille n de X. De là, on peut écrire que

$$P\left[\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right] = 1-\alpha \tag{5.80}$$

où, si $Y \sim \chi^2_{n-1}$ (voir la figure 5.7),

$$P[Y \leq \chi^2_{\alpha/2,n-1}] = 1 - \frac{\alpha}{2} \tag{5.81}$$

et

$$P[Y \leq \chi^2_{1-\alpha/2,n-1}] = 1 - \left(1 - \frac{\alpha}{2}\right) = \frac{\alpha}{2} \tag{5.82}$$

Il s'ensuit qu'un intervalle de confiance à $100(1-\alpha)$ % pour σ^2 est

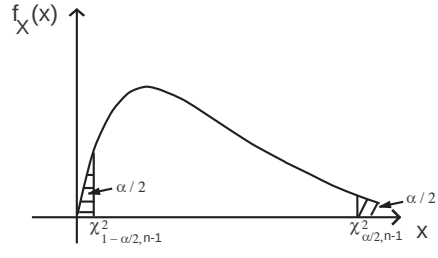

Fig. 5.7. Définition des quantités $\chi^2_{\alpha/2,n-1}$ et $\chi^2_{1-\alpha/2,n-1}$ pour une variable aléatoire X qui présente une distribution χ^2_{n-1}

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] \tag{5.83}$$

Remarques.
(i) L'intervalle de confiance n'est pas symétrique, comme dans le cas des moyennes, puisque la distribution du khi-deux ne l'est pas.

(ii) Comme dans le cas des quantités z_α et $t_{\alpha,n}$, les valeurs de $\chi^2_{\alpha,n}$ sont obtenues en utilisant une calculatrice ou trouvées dans une table. Ces valeurs sont données dans le tableau 5.3, à la page 240, pour différents α et n.

Tableau 5.3. Valeurs de $\chi^2_{\alpha,n}$

n	1	2	3	4	5	6	7	8	9
$\chi^2_{0,005;n}$	7,88	10,60	12,84	14,86	16,75	18,55	20,28	21,96	23,59
$\chi^2_{0,01;n}$	6,63	9,21	11,34	13,28	15,09	16,81	18,48	20,09	21,67
$\chi^2_{0,025;n}$	5,02	7,38	9,35	11,14	12,83	14,45	16,01	17,53	19,02
$\chi^2_{0,05;n}$	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92
$\chi^2_{0,10;n}$	2,71	4,61	6,25	7,78	9,24	10,65	12,02	13,36	14,68
$\chi^2_{0,90;n}$	0,02	0,21	0,58	1,06	1,61	2,20	2,83	3,49	4,17
$\chi^2_{0,95;n}$	0^+	0,10	0,35	0,71	1,15	1,64	2,17	2,73	3,33
$\chi^2_{0,975;n}$	0^+	0,05	0,22	0,48	0,83	1,24	1,69	2,18	2,70
$\chi^2_{0,99;n}$	0^+	0,02	0,11	0,30	0,55	0,87	1,24	1,65	2,09
$\chi^2_{0,995;n}$	0^+	0,01	0,07	0,21	0,41	0,68	0,99	1,34	1,73

n	10	15	20	25	30	40	50	100
$\chi^2_{0,005;n}$	25,19	32,80	40,00	46,93	53,67	66,77	79,49	140,17
$\chi^2_{0,01;n}$	23,21	30,58	37,57	44,31	50,89	63,69	76,15	135,81
$\chi^2_{0,025;n}$	20,48	27,49	34,17	40,65	46,98	59,34	71,42	129,56
$\chi^2_{0,05;n}$	18,31	25,00	31,41	37,65	43,77	55,76	67,50	124,34
$\chi^2_{0,10;n}$	15,99	22,31	28,41	34,28	40,26	51,81	63,17	118,50
$\chi^2_{0,90;n}$	4,87	8,55	12,44	16,47	20,60	29,05	37,69	82,36
$\chi^2_{0,95;n}$	3,94	7,26	10,85	14,61	18,49	26,51	34,76	77,93
$\chi^2_{0,975;n}$	3,25	6,27	9,59	13,12	16,79	24,43	32,36	74,22
$\chi^2_{0,99;n}$	2,56	5,23	8,26	11,52	14,95	22,16	29,71	70,06
$\chi^2_{0,995;n}$	2,16	4,60	7,43	10,52	13,79	20,71	27,99	67,33

Il existe aussi des formules qui donnent des approximations excellentes, en particulier:

$$\chi^2_{\alpha,n} \simeq n\left[z_\alpha\sqrt{\frac{2}{9n}} + 1 - \frac{2}{9n}\right]^3 \quad \text{(approximation de Wilson-Hilferty)} \qquad (5.84)$$

et

$$\chi^2_{\alpha,n} \simeq \frac{1}{2}\left[z_\alpha + \sqrt{2n-1}\right]^2 \quad \text{(approximation de Fisher)} \qquad (5.85)$$

De plus, si n est assez grand, alors on peut utiliser l'approximation $\chi^2_n \approx \mathrm{N}(n, 2n)$ (obtenue par le théorème central limite).

(b) **Cas où μ est connu**

Si la moyenne μ de la population est connue, alors un intervalle de confiance plus précis pour la variance σ^2 est obtenu à partir de la statistique

$$\hat{\sigma}^2 := \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n} \tag{5.86}$$

Par définition, on a:

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n \tag{5.87}$$

Alors l'intervalle de confiance à $100(1-\alpha)$ % pour σ^2 devient

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{\alpha/2,n}}, \frac{n\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n}} \right] \tag{5.88}$$

(c) **Cas où X ne présente pas nécessairement une distribution normale**

Si X ne présente pas une distribution normale (ou bien si on ne peut pas l'affirmer), on peut tout de même donner une formule pour un intervalle de confiance approximatif pour l'écart-type σ de la population, pourvu que n soit suffisamment grand. En effet, si $n \geq 30$, alors on peut écrire que

$$S \approx \mathrm{N}\left(\sigma, \tfrac{\sigma^2}{2n}\right) \tag{5.89}$$

De là, on trouve qu'un intervalle de confiance approximatif à $100(1-\alpha)$ % pour σ est donné par

$$\left[\frac{S}{1 + \frac{z_{\alpha/2}}{\sqrt{2n}}}, \frac{S}{1 - \frac{z_{\alpha/2}}{\sqrt{2n}}} \right] \tag{5.90}$$

Remarques.
(i) En élevant les deux bornes au carré dans l'intervalle précédent, on obtient un intervalle de confiance approximatif pour σ^2.

(ii) On suppose ci-dessus que μ est inconnu. Si l'on connaît la moyenne de la population, on devrait en tenir compte pour obtenir un intervalle de confiance plus précis.

5.4.5 Intervalle de confiance pour σ_X^2/σ_Y^2

Soit $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ des échantillons aléatoires des variables aléatoires *indépendantes* $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$, respectivement. On suppose que tous les paramètres sont inconnus.

Puisque $(m-1)S_X^2/\sigma_X^2 \sim \chi_{m-1}^2$ et $(n-1)S_Y^2/\sigma_Y^2 \sim \chi_{n-1}^2$ sont des variables aléatoires *indépendantes*, on peut écrire que

$$\frac{S_Y^2/\sigma_Y^2}{S_X^2/\sigma_X^2} \sim F_{n-1,m-1} \tag{5.91}$$

Alors un intervalle de confiance à $100(1-\alpha)$ % pour le quotient σ_X^2/σ_Y^2 est

$$\left[\frac{S_X^2}{S_Y^2} F_{1-\alpha/2,n-1,m-1}, \frac{S_X^2}{S_Y^2} F_{\alpha/2,n-1,m-1}\right] \tag{5.92}$$

où la quantité $F_{\alpha/2,n_1,n_2}$, pour n_1 et $n_2 \in \{1, 2, \ldots\}$, est définie comme suit (voir la figure 5.8):

$$P[F \leq F_{\alpha/2,n_1,n_2}] = 1 - \frac{\alpha}{2} \quad \text{si } F \sim F_{n_1,n_2} \tag{5.93}$$

De même,

$$P[F \leq F_{1-\alpha/2,n_1,n_2}] = 1 - \left(1 - \frac{\alpha}{2}\right) = \frac{\alpha}{2} \tag{5.94}$$

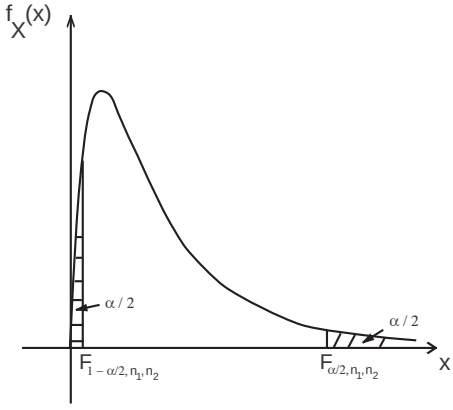

Fig. 5.8. Définition des quantités $F_{\alpha/2,n_1,n_2}$ et $F_{1-\alpha/2,n_1,n_2}$ pour une variable aléatoire X qui présente une distribution F_{n_1,n_2}

Remarque. Encore une fois, on trouve les F_{α,n_1,n_2} à l'aide d'une calculatrice ou dans une table statistique. Quelques valeurs de F_{α,n_1,n_2}, pour les α les plus importants, sont données dans le tableau 5.4, à la page 243, dans le cas où $n_1 = n_2 := n$. De plus, on a la relation suivante:

$$F_{1-\alpha,n_1,n_2} = \frac{1}{F_{\alpha,n_2,n_1}} \tag{5.95}$$

Tableau 5.4. Valeurs de F_{α,n_1,n_2}

n	1	2	3	4	5	6	7	8
$F_{0,01;n,n}$	4052	99,00	29,46	15,98	10,97	8,47	6,99	6,03
$F_{0,025;n,n}$	647,8	39,00	15,44	9,60	7,15	5,82	4,99	4,43
$F_{0,05;n,n}$	161,4	19,00	9,28	6,39	5,05	4,28	3,79	3,44
$F_{0,10;n,n}$	39,86	9,00	5,39	4,11	3,45	3,05	2,78	2,59

n	9	10	15	20	30	40	50	100	200
$F_{0,01;n,n}$	5,35	4,85	3,52	2,94	2,39	2,11	1,95	1,60	1,39
$F_{0,025;n,n}$	4,03	3,72	2,86	2,46	2,07	1,88	1,75	1,48	1,32
$F_{0,05;n,n}$	3,18	2,98	2,40	2,12	1,84	1,69	1,60	1,39	1,26
$F_{0,10;n,n}$	2,44	2,32	1,97	1,79	1,61	1,51	1,44	1,29	1,20

Les valeurs de $F_{0,025;n_1,n_2}$ et $F_{0,05;n_1,n_2}$ sont données à l'appendice B, à la page 519, pour n_1 et $n_2 \in \{1, 2, \ldots, 12\}$. Finalement, on peut utiliser la formule d'approximation qui suit pour des valeurs de n_1 et n_2 assez grandes:

$$F_{\alpha,n_1,n_2} \simeq \exp\left[\frac{1}{n_2} - \frac{1}{n_1} + z_\alpha\sqrt{\frac{2}{n_1} + \frac{2}{n_2}}\right] \tag{5.96}$$

5.4.6 Intervalle de confiance pour p

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X qui présente une distribution de Bernoulli de paramètre p; c'est-à-dire que $X \sim \mathrm{B}(1, p)$. On peut écrire que

$$\sum_{i=1}^{n} X_i \sim \mathrm{B}(n, p) \tag{5.97}$$

Puisque $\mathrm{B}(n, p) \approx \mathrm{N}(np, np(1-p))$, si n est assez grand (voir la page 90), on a:

$$\bar{X} \approx \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) \tag{5.98}$$

Ainsi,

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right] \simeq 1-\alpha \tag{5.99}$$

Maintenant, pour obtenir une formule pour un intervalle de confiance approximatif à $100(1-\alpha)$ % pour le paramètre inconnu p, on le remplace par son estimateur $\hat{p} = \bar{X}$ dans la racine carrée ci-dessus. On obtient alors l'intervalle suivant:

$$\left[\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right] \tag{5.100}$$

Remarque. Lorsqu'on calcule un intervalle de confiance particulier, il faut que la limite inférieure de confiance obtenue soit supérieure ou égale à 0; de même, la limite supérieure de confiance doit être inférieure ou égale à 1. Sinon, on remplace la limite inférieure par 0 et la limite supérieure par 1.

Exemple 5.4.3. Supposons que, dans un sondage réalisé avant une élection, 35 % des 1000 personnes interrogées ont exprimé leur préférence pour le candidat A. Alors un intervalle de confiance (approximatif) à 95 %, pour le pourcentage de votes que ce candidat recueillera lors de l'élection, est donné par

$$0{,}35 \pm (1{,}960)\sqrt{\frac{(0{,}35)(0{,}65)}{1000}} \simeq 0{,}35 \pm 0{,}03$$

Dans le langage courant (à la télévision, par exemple), on dit que la marge d'erreur du sondage, selon lequel la popularité du candidat A est de 35 % des électeurs, est de plus ou moins 3 %, et ce, *19 fois sur 20.* ♢

Si l'on désire que l'erreur maximale (avec un coefficient de confiance de $100(1-\alpha)$ %) que l'on commet en estimant p par la moyenne $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ soit égale à une constante donnée $E_{\max}$, alors il faut prendre environ

$$n = \left(\frac{z_{\alpha/2}}{E_{\max}}\right)^2 \hat{p}(1-\hat{p}) \tag{5.101}$$

observations. Cependant, $\hat{p}$ est inconnu tant que l'on n'a pas recueilli les données. Par conséquent, pour être prudent, on remplace le produit $\hat{p}(1-\hat{p})$ par la plus

grande valeur qu'il peut prendre, soit (0,5)(1 - 0,5) = 0,25 (car $\hat{p} \in [0, 1]$). Donc, la formule précédente devient

$$n = \frac{1}{4}\left(\frac{z_{\alpha/2}}{E_{\max}}\right)^2 \tag{5.102}$$

ce qui correspond au *pire cas*. En général, on doit arrondir le nombre obtenu à l'entier supérieur à ce nombre. Avec cette valeur de n, le coefficient de confiance est *au moins* égal à $100(1-\alpha)$ %.

Si l'on a déjà recueilli un certain nombre (assez grand) de données, alors on peut se servir de l'estimation ponctuelle de p, calculée à partir de ces données, dans la formule (5.101).

5.4.7 Intervalle de confiance pour $p_X - p_Y$

Soit $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ des échantillons aléatoires de variables aléatoires *indépendantes* $X \sim \mathrm{B}(1, p_X)$ et $Y \sim \mathrm{B}(1, p_Y)$, respectivement. Puisque

$$\frac{\bar{X} - \bar{Y} - (p_X - p_Y)}{\left[\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}\right]^{1/2}} \approx \mathrm{N}(0, 1) \tag{5.103}$$

un intervalle de confiance approximatif à $100(1-\alpha)$ % pour la différence $p_X - p_Y$ est (sous forme compacte)

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\left[\frac{\bar{X}(1-\bar{X})}{m} + \frac{\bar{Y}(1-\bar{Y})}{n}\right]^{1/2} \tag{5.104}$$

5.4.8 Intervalle de confiance basé sur θ_{VM}

Soit $\hat{\theta} = \theta_{VM}$, l'estimateur à vraisemblance maximale du paramètre inconnu θ. Si n est assez grand, alors on peut écrire que

$$\hat{\theta} \approx \mathrm{N}\left(\theta, \frac{1}{nB^2}\right) \tag{5.105}$$

où

$$B^2 := E\left[\frac{d}{d\theta} \ln f_X(x; \theta)\right]^2 \tag{5.106}$$

Il s'ensuit qu'un intervalle de confiance approximatif à $100(1-\alpha)$ % pour θ est

$$\left[\hat{\theta} - z_{\alpha/2}\frac{1}{\sqrt{n}B}, \hat{\theta} + z_{\alpha/2}\frac{1}{\sqrt{n}B}\right] \tag{5.107}$$

Remarque. Si B contient le paramètre inconnu θ, alors on le remplace par $\hat{\theta}$.

Exemple 5.4.4. Soit X une variable aléatoire qui présente une distribution de Poisson de paramètre θ. On a déjà trouvé (voir l'exemple 5.2.5, page 220) que $B^2 = 1/\theta$ dans ce cas. Puisque l'on a aussi trouvé (voir l'exemple 5.2.7b, page 222) que $\theta_{VM} = \bar{X}$, on peut écrire qu'un intervalle de confiance approximatif à $100(1-\alpha)$ % pour θ est donné par

$$\left[\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}\right]$$

Notons que l'on obtient le même résultat ici en utilisant le fait que $\bar{X}$ présente approximativement une distribution $\mathrm{N}\big(\theta, \frac{\theta}{n}\big)$, par le théorème central limite, et en remplaçant θ par son estimateur $\bar{X}$ dans la variance. ♢

5.5 Exercices du chapitre 5

Exercices résolus

Question n$^{\text{o}}$ 1

Soit X_1, X_2, …, X_{20} un échantillon aléatoire d'une distribution $\mathrm{N}(0,1)$. On pose:

$$\bar{X}_1 = \frac{1}{4}\sum_{i=1}^{4} X_i, \quad \bar{X}_2 = \frac{1}{16}\sum_{i=5}^{20} X_i$$

$$S_1^2 = \frac{1}{3}\sum_{i=1}^{4}\big(X_i - \bar{X}_1\big)^2, \quad S_2^2 = \frac{1}{15}\sum_{i=5}^{20}\big(X_i - \bar{X}_2\big)^2$$

(a) Calculer $P\left[-1{,}638 \le \frac{2\bar{X}_1}{S_1}\right]$.

(b) Calculer $P\left[\frac{2\bar{X}_1}{S_2} < 2{,}602\right]$.

Solution. (a) On a: $\sqrt{4}\bar{X}_1/S_1 \sim t_3$. Or, on trouve que $1{,}638 \overset{\text{tab. } 5.2}{\simeq} t_{0,10;3}$. Il s'ensuit que $-1{,}638 \simeq t_{0,90;3}$. Alors

$$P\left[-1{,}638 \leq \frac{2\bar{X}_1}{S_1}\right] \simeq 0{,}90$$

(b) D'abord, on a: $\sqrt{4}\bar{X}_1/S_2 \sim t_{15}$ et $2{,}602 \overset{\text{tab. 5.2}}{\simeq} t_{0{,}01;15}$. Cela implique que

$$P\left[\frac{2\bar{X}_1}{S_2} < 2{,}602\right] \simeq 1 - 0{,}01 = 0{,}99$$

Question n° 2

Supposons que $X \sim \chi^2_1$ et $Y \sim \mathrm{G}(\alpha = 3/2, \lambda = 1/2)$ sont deux variables aléatoires indépendantes. Quelle distribution présente $Z := X + Y$? Donner aussi le ou les paramètres de cette distribution.

Solution. On peut écrire que $Y \sim \mathrm{G}(\alpha = 3/2, \lambda = 1/2) \equiv \chi^2_3$. Alors, par indépendance, il s'ensuit que $Z = X + Y \sim \chi^2_4$ ($\equiv \mathrm{G}(\alpha = 2, \lambda = 1/2)$).

Question n° 3

Soit X_1, X_2 et X_3 des variables aléatoires indépendantes qui présentent toutes les trois une distribution χ^2_1. Trouver la constante c si

$$Y := \frac{2X_1 + cX_2}{X_2 + X_3}$$

présente une distribution de Fisher. Donner aussi la distribution précise de la variable aléatoire Y.

Solution. Pour que $\frac{2X_1+cX_2}{X_2+X_3}$ présente une distribution de Fisher, il faut que $c = 0$, sinon le numérateur et le dénominateur ne sont pas indépendants. On a alors:

$$Y = \frac{2X_1}{X_2 + X_3} \sim F_{1,2}$$

Question n° 4

Soit X une variable aléatoire $\mathrm{N}(0, 1)$ et Y une variable aléatoire $\mathrm{N}(0, 4)$. On suppose que X et Y sont indépendantes. Calculer $P\left[(X + Y)^2 < 25{,}1\right]$.

Solution. On a: $X + Y \overset{\text{ind.}}{\sim} \mathrm{N}(0 + 0, 1 + 4) \equiv \mathrm{N}(0, 5) \Rightarrow \left(\frac{X+Y-0}{\sqrt{5}}\right)^2 \sim \chi^2_1$. Alors

$$P[(X + Y)^2 < 25{,}1] = P[\chi^2_1 < 5{,}02] \overset{\text{tab. 5.3}}{\simeq} 1 - 0{,}025 = 0{,}975$$

Remarque. On peut aussi écrire que

$$P[(X+Y)^2 < 25{,}1] \simeq P\left[|\mathrm{N}(0,1)| < \frac{\sqrt{25{,}1}}{\sqrt{5}}\right]$$
$$\simeq 2\,\Phi(2{,}24) - 1 \overset{\text{tab. A.3}}{\simeq} 2(0{,}9875) - 1 = 0{,}975$$

Question n° 5

On a recueilli 10 observations particulières, $x_1, \ldots, x_{10}$, d'une variable aléatoire discrète X. Les sommes suivantes ont été calculées:

$$\sum_{i=1}^{10} x_i = 0, \quad \sum_{i=1}^{10} x_i^2 = 18, \quad \sum_{i=1}^{10} x_i^3 = 0, \quad \sum_{i=1}^{10} x_i^4 = 110$$

En se basant sur les quantités $\hat{\beta}_1$ et $\hat{\beta}_2$, peut-on dire que l'on peut approcher la distribution de X par une distribution normale? Justifier votre réponse.

Solution. On a: $\bar{x} = 0$; alors $s^2 = \frac{1}{9}\sum_{i=1}^{10} x_i^2 = 2$, $\hat{\mu}_3 = \frac{1}{9}\sum_{i=1}^{10} x_i^3 = 0$ et $\hat{\mu}_4 = \frac{1}{9}\sum_{i=1}^{10} x_i^4 = 110/9$. De là, on calcule

$$\hat{\beta}_1 = 0 \quad \text{et} \quad \hat{\beta}_2 = \frac{\hat{\mu}_4}{s^4} = \frac{110}{36} \simeq 3{,}06$$

Puisque $\beta_1 = 0$ et $\beta_2 = 3$ dans le cas d'une distribution normale, l'approximation semble bonne. Toutefois, $n = 10$ est petit.

Question n° 6

Supposons que

x	0	1	2	3
$p_X(x)$	1/4	1/4	1/4	1/4

Un échantillon aléatoire sans remise de taille 2 est tiré de la population de taille $N = 4$. On définit $\bar{X} = (X_1 + X_2)/2$. Calculer $\mu_{\bar{X}}$ et $\sigma_{\bar{X}}$.

Solution. On trouve facilement que $\mu_X = 3/2$ et $\sigma_X^2 = 5/4$. Il s'ensuit que

$$\mu_{\bar{X}} = \mu_X = \frac{3}{2} \quad \text{et} \quad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{2}}\left(\frac{4-2}{4-1}\right)^{1/2} = (5/12)^{1/2} \simeq 0{,}6455$$

Question n° 7

Un autobus passe à un certain coin de rue tous les matins vers 9 h. Soit X la différence (en minutes) entre l'instant où l'autobus passe et 9 h. On suppose que

X présente (approximativement) une distribution normale $\mathrm{N}\big(\mu = 0, \sigma^2 = 25\big)$. On considère deux journées indépendantes. Soit X_k la valeur de la variable aléatoire X lors de la k^{e} journée, pour $k = 1, 2$. Trouver le nombre c tel que $P\left[X_1^2 + X_2^2 < c\right] = 0{,}95$.

Solution. On a:

$$P[X_1^2 + X_2^2 < c] = P\left[\frac{X_1^2 + X_2^2}{25} < \frac{c}{25}\right] = P\left[\chi_2^2 < \frac{c}{25}\right] = 0{,}95$$
$$\iff \frac{c}{25} \overset{\text{tab. } 5.3}{\simeq} 5{,}99 \iff c \simeq 149{,}75$$

Question n$^{\mathrm{o}}$ 8

Des plaques de laiton sont soumises à un contrôle visuel. La surface d'une plaque est vérifiée pour détecter des taches de cuivre ou d'oxygénation ou d'autres défectuosités apparentes. On a noté le nombre de défauts par plaque sur 50 plaques:

Nombre de défauts	0	1	2	3	4
Nombre de plaques	10	15	12	8	5

Calculer (a) l'écart-type et (b) la médiane de l'échantillon.

Solution. (a) On a:

$$\bar{x} = \frac{1}{50}(10 \times 0 + \ldots + 5 \times 4) = \frac{83}{50} = 1{,}66$$

de sorte que

$$s = \left\{\frac{1}{49}[10(0 - 1{,}66)^2 + \ldots + 5(4 - 1{,}66)^2]\right\}^{1/2} \simeq 1{,}255$$

(b) Puisqu'il y a 50 données, la médiane est égale à la moyenne arithmétique des 25^{e} et 26^{e} données placées en ordre croissant. Il s'ensuit que la médiane est égale à $\frac{1+2}{2} = 1{,}5$.

Question n$^{\mathrm{o}}$ 9

(a) Quelle est la probabilité que la variance d'un échantillon aléatoire de taille $n = 16$, tiré d'une population normale de moyenne nulle et de variance égale à 3, soit inférieure à 5?

(b) Supposons que σ est inconnu en (a). Si l'écart-type de l'*échantillon* est égal à 4, quelle est (environ) la probabilité que la moyenne de l'échantillon aléatoire soit supérieure à 1,75?

Solution. (a) On a:

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \implies 15\frac{S^2}{3} \sim \chi^2_{15}$$

On cherche

$$P[S^2 < 5] = P[\chi^2_{15} < 25] = 1 - P[\chi^2_{15} \geq 25] \overset{\text{tab. 5.3}}{\simeq} 1 - 0{,}05 = 0{,}95$$

(b) On sait que

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1)$$

Puisque σ est inconnu, on le remplace par $s = 4$, et l'on écrit que

$$P[\bar{X} > 1{,}75] \simeq P\left[\mathrm{N}(0,1) > \frac{1{,}75 - 0}{4/\sqrt{16}}\right] = 1 - \Phi(1{,}75) \overset{\text{tab. A.3}}{\simeq} 1 - 0{,}96 = 0{,}04$$

Remarque. On a vu que $(\bar{X} - \mu)/(S/\sqrt{n})$ présente une distribution t_{n-1}. Cependant, on ne peut pas écrire que $(\bar{X} - \mu)/(s/\sqrt{n}) \approx t_{n-1}$, car s est une *constante*, et non pas une variable aléatoire comme S.

Question n$^{\text{o}}$ 10

Parmi 100 personnes prises au hasard dans un groupe de patients souffrant d'un cancer du poumon, 67 sont mortes moins de 5 ans après la détection de la maladie.

(a) Calculer un intervalle de confiance à 95 % pour la probabilité qu'une personne, dont le cancer du poumon vient d'être détecté, meure d'ici 5 ans.

(b) Quel est le nombre minimal d'observations supplémentaires dont on a besoin, dans le pire des cas, pour pouvoir estimer la probabilité en (a) avec une erreur maximale de 0,02, et ce, avec un coefficient de confiance de 95 %?

Solution. (a) L'intervalle de confiance est donné par

$$\hat{p} \pm z_{0{,}05/2}\left[\frac{\hat{p}(1-\hat{p})}{100}\right]^{1/2}$$

où $\hat{p} = \bar{x} = \frac{67}{100}$. Étant donné que $z_{0{,}025} \simeq 1{,}960$, on obtient (environ) l'intervalle suivant: $0{,}670 \pm 0{,}092 \Leftrightarrow [0{,}578; 0{,}762]$.

(b) On veut que

$$z_{0,05/2}\left[\frac{\hat{p}(1-\hat{p})}{n}\right]^{1/2} \leq 0{,}02$$

où n est le nombre *total* d'observations dont on a besoin. Puisque $\hat{p} = 1/2$ dans le pire des cas, on peut écrire que

$$n \geq 0{,}25\left(\frac{z_{0,025}}{0{,}02}\right)^2 \simeq 2401$$

Donc, il faut au moins 2301 observations supplémentaires.

Question n$^{\text{o}}$ 11

Vingt-cinq nombres aléatoires provenant d'une distribution $\mathrm{N}(0,1)$ sont générés par un ordinateur. On suppose que ces nombres sont indépendants. Quelle est la probabilité que

(a) la somme des nombres obtenus soit inférieure à 5?

(b) la somme des carrés des nombres obtenus soit située dans l'intervalle $[13{,}12; 14{,}61]$?

(c) le résultat de la division du premier nombre par la racine carrée de la somme des carrés des 20 derniers soit supérieur à 0,4664?

(d) le résultat de la division de la somme des carrés des cinq premiers nombres par la somme des carrés des cinq derniers soit inférieur à 5,05?

Solution. (a) Soit $X_1, X_2, \ldots, X_{25}$ les nombres générés. Par indépendance, on peut écrire que $T := \sum_{i=1}^{25} X_i \sim \mathrm{N}(0, 25)$. On cherche

$$P[T < 5] = \Phi\left(\frac{5-0}{5}\right) = \Phi(1) \overset{\text{tab. A.3}}{\simeq} 0{,}8413$$

(b) Par définition, on a: $Y := \sum_{i=1}^{25} X_i^2 \sim \chi^2_{25}$. On calcule alors

$$P[13{,}12 \leq Y \leq 14{,}61] \simeq 0{,}975 - 0{,}95 = 0{,}025$$

car (voir le tableau 5.3, page 240) $\chi^2_{0,975;25} \simeq 13{,}12$ et $\chi^2_{0,95;25} \simeq 14{,}61$.

(c) On a:

$$\frac{X_1}{\left(\sum_{i=6}^{25} X_i^2/20\right)^{1/2}} \sim t_{20}$$

Il s'ensuit que

$$P\left[\frac{X_1}{\left(\sum_{i=6}^{25} X_i^2\right)^{1/2}} > 0{,}4664\right] \simeq P[t_{20} > 2{,}086] \overset{\text{tab. } 5.2}{\simeq} 0{,}025$$

(d) On peut écrire que

$$\frac{X_1^2 + \ldots + X_5^2}{X_{21}^2 + \ldots + X_{25}^2} \sim F_{5,5}$$

Alors

$$P\left[\frac{X_1^2 + \ldots + X_5^2}{X_{21}^2 + \ldots + X_{25}^2} < 5{,}05\right] = P[F_{5,5} < 5{,}05] \overset{\text{tab. } 5.4}{\simeq} 1 - 0{,}05 = 0{,}95$$

Question n° 12

Soit $x_1, \ldots, x_n$ n nombres réels. On définit

$$y_i = \frac{x_i - \bar{x}}{s} \quad \text{pour } i = 1, \ldots, n$$

où $\bar{x}$ est la moyenne et s est l'écart-type des n nombres.

(a) Calculer la moyenne $\bar{y}$ et la variance s_y^2 de $y_1, \ldots, y_n$.

(b) Calculer la moyenne de $y_1^2, \ldots, y_n^2$.

Solution. (a) On a:

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s} = 0$$

Alors on peut écrire que

$$\begin{aligned} s_y^2 &= \sum_{i=1}^{n} \frac{y_i^2}{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{s^2} \\ &= \frac{1}{s^2} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{s^2}{s^2} = 1 \end{aligned}$$

(b) On calcule

$$\sum_{i=1}^{n} \frac{y_i^2}{n} = \frac{n-1}{n} \sum_{i=1}^{n} \frac{y_i^2}{n-1} \overset{\text{(a)}}{=} \frac{n-1}{n} \cdot 1 = \frac{n-1}{n}$$

Question n° 13

On considère une population $\mathrm{N}(0,1)$. On y prélève trois échantillons aléatoires indépendants de taille $n = 6$. Soit S_1^2, S_2^2 et S_3^2 les variances des échantillons aléatoires. Calculer

$$P\left[\sum_{i=1}^{3} S_i^2 > 5\right]$$

Solution. On peut écrire que $(6-1)S_i^2/1 \sim \chi_5^2$ pour $i = 1, 2, 3$. Alors, par indépendance, on peut affirmer que $5\sum_{i=1}^{3} S_i^2 \sim \chi_{15}^2$. Il s'ensuit que

$$P\left[\sum_{i=1}^{3} S_i^2 > 5\right] = P[\chi_{15}^2 > 25]$$

Or, on trouve dans la tableau 5.3, à la page 240, que $25{,}0 \simeq \chi_{0{,}05;15}^2$. Donc, la probabilité recherchée est environ égale à 0,05.

Question n° 14

Soit $x_1, \ldots, x_n$ n observations particulières d'une variable aléatoire X.

(a) On pose $S(c) = \sum_{i=1}^{n} (x_i - c)^2$. La valeur de c qui minimise la fonction $S(c)$ est-elle la moyenne $\bar{x}$ ou la médiane $\tilde{x}$ des observations? Justifier votre réponse.

(b) Le résultat $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})\, x_i$ est-il exact? Justifier votre réponse.

Solution. (a) On a:

$$\frac{d}{dc} S(c) = 2 \sum_{i=1}^{n} (x_i - c)(-1) = 0 \iff \sum_{i=1}^{n} x_i - nc = 0 \iff c = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}$$

De plus,

$$\frac{d^2}{dc^2} S(c) = -2 \sum_{i=1}^{n} (-1) = 2n > 0$$

Donc, c'est $\bar{x}$ qui minimise la fonction $S(c)$.

(b) On peut écrire que

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})x_i-\bar{x}\sum_{i=1}^{n}(x_i-\bar{x})$$
$$=\sum_{i=1}^{n}(x_i-\bar{x})x_i \quad (\text{car } \textstyle\sum_{i=1}^{n}(x_i-\bar{x})=0)$$

Donc, le résultat est exact.

Question n$^{\text{o}}$ 15

On prend 100 nombres au hasard dans l'intervalle [0,1]. Si l'on veut calculer la probabilité que la somme des carrés des 100 nombres soit dans un intervalle donné, doit-on utiliser la distribution normale ou celle du khi-deux? Justifier votre réponse.

Solution. On doit utiliser la distribution normale, car la distribution de la somme $\sum_{k=1}^{100} X_k^2$, où X_k désigne le k^{e} nombre pris au hasard, est approximativement normale, selon le théorème central limite.

Remarque. La distribution *exacte* de la somme des carrés serait celle du khi-deux avec 100 degrés de liberté si les X_k étaient des variables aléatoires $\mathrm{N}(0,1)$ (indépendantes), plutôt que des variables $\mathrm{U}[0,1]$.

Question n$^{\text{o}}$ 16

Soit

$$p_X(x;\theta)=(\theta-1)^{x-1}\,\theta^{-x} \quad \text{pour } x=1,2,\ldots$$

où $\theta>1$. On peut montrer que $E[X]=\theta$ et $\mathrm{VAR}[X]=\theta(\theta-1)$.

(a) Déterminer l'estimateur θ_M de θ par la méthode des moments.

(b) Calculer l'erreur quadratique moyenne de θ_M.

(c) Obtenir un intervalle de confiance approximatif à 95 % pour θ si un échantillon aléatoire de X, de taille $n=50$, a donné une moyenne $\bar{x}$ égale à 2.

Solution. (a) On pose:

$$E[X]=\bar{X} \iff \theta=\bar{X}$$

Donc, l'estimateur de θ par la méthode des moments est $\theta_M=\bar{X}$.

(b) On a: $E[\theta_M] \overset{(a)}{=} E\left[\bar{X}\right] = E[X] = \theta$. Alors on peut écrire que

$$\text{E.Q.M.}[\theta_M] = \text{VAR}[\theta_M] = \text{VAR}[\bar{X}] = \frac{\text{VAR}[X]}{n} = \frac{\theta(\theta-1)}{n}$$

(c) On déduit du théorème central limite et de la partie (b) que

$$\theta_M = \bar{X} \approx \text{N}\left(\theta, \frac{\theta(\theta-1)}{n}\right)$$

Il s'ensuit que

$$P\left[-z_{\alpha/2} \leq \frac{\theta_M - \theta}{\sqrt{\frac{\theta(\theta-1)}{n}}} \leq z_{\alpha/2}\right] \simeq 1-\alpha$$

$$\Longrightarrow P\left[-z_{\alpha/2} \leq \frac{\theta_M - \theta}{\sqrt{\frac{\theta_M(\theta_M-1)}{n}}} \leq z_{\alpha/2}\right] \simeq 1-\alpha$$

De là, on trouve que l'intervalle de confiance approximatif à $100(1-\alpha)$ % pour θ est donné par

$$\theta_M \pm z_{\alpha/2}\sqrt{\frac{\theta_M(\theta_M-1)}{n}} \iff \bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}\left(\bar{X}-1\right)}{n}}$$

Avec $\alpha = 0{,}05$, $n = 50$ et $\bar{x} = 2$, on obtient:

$$2 \pm \underbrace{z_{0,025}}_{1{,}960}\sqrt{\frac{2(2-1)}{50}} \iff 2 \pm 0{,}39 \quad \text{(environ)}$$

Question n$^{\text{o}}$ 17

On suppose que le temps de service (en minutes) à un comptoir est une variable aléatoire X dont la fonction de densité est donnée par

$$f_X(x;\theta) = 2\theta^2 x^2 e^{-\theta x^2} \quad \text{pour } x \geq 0$$

où $\theta > 0$. Trouver l'estimateur à vraisemblance maximale de θ.

Solution. Soit $X_1, \ldots, X_n$ un échantillon aléatoire de X. On pose:

$$L(\theta) = \prod_{i=1}^{n} f_X(X_i;\theta) = \prod_{i=1}^{n} 2\theta^2 X_i^2 \exp\left(-\theta X_i^2\right) \quad (\text{si } X_i \geq 0 \ \forall i)$$
$$= 2^n \theta^{2n} \left(\prod_{i=1}^{n} X_i^2\right) \exp\left(-\theta \sum_{i=1}^{n} X_i^2\right)$$

Alors

$$\ln L(\theta) = n \ln 2 + 2n \ln \theta + 2 \sum_{i=1}^{n} \ln X_i - \theta \sum_{i=1}^{n} X_i^2$$

et

$$\frac{d}{d\theta} \ln L(\theta) = \frac{2n}{\theta} - \sum_{i=1}^{n} X_i^2 = 0 \iff \theta = \frac{2n}{\sum_{i=1}^{n} X_i^2}$$
$$\Longrightarrow \quad \theta_{VM} = \frac{2n}{\sum_{i=1}^{n} X_i^2}$$

Remarque. On a: $\frac{d^2}{d\theta^2} \ln L(\theta) = -\frac{2n}{\theta^2} < 0$.

Question n° 18

On considère les observations particulières suivantes d'une variable aléatoire: 2, 5, 1, 3, −2 et 6.

(a) Quelle est la médiane des observations?

(b) Calculer l'écart-type de la moyenne des observations.

Solution. (a) On a:

$$\tilde{x} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{2+3}{2} = 2{,}5$$

(b) On trouve que $s \simeq 2{,}88$; alors l'écart-type de la moyenne est $s/\sqrt{6} \simeq 1{,}18$.

Question n° 19

Soit $(x_1, y_1), \ldots, (x_{10}, y_{10})$ 10 observations particulières d'un couple de variables aléatoires (X, Y). On a $\sum_{i=1}^{10} x_i = \sum_{i=1}^{10} y_i = 55$, $s_x = s_y = 3{,}03$ et $\sum_{i=1}^{10} x_i y_i = 380$.

(a) Calculer le coefficient de corrélation $r_{x,y}$.

(b) Calculer $\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2$.

Solution. (a) On a:

$$s_{x,y} = \frac{1}{9}\left(\sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y}\right) = \frac{1}{9}\left(380 - 10\,\frac{55}{10}\,\frac{55}{10}\right) \simeq 8{,}61$$

$$\Longrightarrow \qquad r_{x,y} = \frac{s_{x,y}}{s_x s_y} \simeq 0{,}94$$

(b) On calcule

$$\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 = \sum_{i=1}^{10} (x_i - \bar{x})^2 = 9\, s_x^2 \simeq 82{,}6$$

Question n$^{\text{o}}$ 20

Soit $X \sim \mathrm{N}\big(0, \sigma^2\big)$. Pour estimer le paramètre inconnu σ^2, on propose d'utiliser $\hat{\sigma}^2 := \sum_{i=1}^{n} a X_i^2$, où $X_1, \ldots, X_n$ est un échantillon aléatoire de taille n de X. On $E\left[\hat{\sigma}^2\right] = na\sigma^2$ et $\mathrm{VAR}\big[\hat{\sigma}^2\big] = 2na^2\sigma^4$.

(a) Calculer le biais de $\hat{\sigma}^2$.

(b) Trouver la constante a telle que l'erreur quadratique moyenne de $\hat{\sigma}^2$ soit minimale.

Solution. (a) On a:

$$\mathrm{Biais}\left[\hat{\sigma}^2\right] = E\left[\hat{\sigma}^2\right] - \sigma^2 = na\sigma^2 - \sigma^2 = (na-1)\,\sigma^2$$

(b) On peut écrire que

$$\text{E.Q.M.}\left[\hat{\sigma}^2\right] \overset{\text{(a)}}{=} 2na^2\sigma^4 + (na-1)^2\,\sigma^4$$

Alors

$$\frac{d}{da}\text{E.Q.M.}\left[\hat{\sigma}^2\right] = 4na\sigma^4 + 2(na-1)n\sigma^4 = 0$$

$$\Longleftrightarrow\ 2a + (na-1) = 0 \ \Longleftrightarrow\ a = \frac{1}{n+2}$$

Remarque. Puisque

$$\frac{d^2}{da^2}\text{E.Q.M.}\left[\hat{\sigma}^2\right] = 4n\sigma^4 + 2n^2\sigma^4 > 0$$

pour toute constante a, on peut affirmer qu'il s'agit bien d'un minimum.

Question n^{o} 21

Soit

$$f_X(x;\theta) = 2x/\theta^2 \quad \text{pour } 0 < x < \theta$$

On cherche à obtenir l'estimateur à vraisemblance maximale, θ_{VM}, de θ.

(a) Calculer la fonction de vraisemblance $L(\theta)$.

(b) Trouver l'estimateur θ_{VM}.

Solution. (a) On a:

$$L(\theta) := \prod_{i=1}^{n} f_X(X_i;\theta) = \prod_{i=1}^{n} 2X_i\theta^{-2} = 2^n\theta^{-2n}\prod_{i=1}^{n} X_i$$

si $0 < X_i < \theta \ \forall i$.

(b) Notons d'abord que, à cause de la condition $0 < X_i < \theta \ \forall i$ ci-dessus, on ne peut pas se servir de la dérivée de la fonction $L(\theta)$ pour trouver θ_{VM}. Cependant, puisque $L(\theta)$ est une fonction strictement *décroissante*, on peut affirmer que

$$\theta_{VM} = \max\{X_1, \ldots, X_n\}$$

C'est-à-dire que θ_{VM} est la plus grande observation dans l'échantillon aléatoire $X_1, \ldots, X_n$.

Question n^{o} 22

Soit X une variable aléatoire qui présente une distribution de Poisson de paramètre θ.

(a) Quelle distribution présente approximativement l'estimateur à vraisemblance maximale de θ? Donner aussi le ou les paramètres de cette distribution.

(b) Quel est l'estimateur à vraisemblance maximale de $\beta := 1/\theta$?

Solution. (a) On a (voir l'exemple 5.2.7 (a), page 222): $\theta_{VM} = \bar{X}$; alors, par le théorème central limite, on peut écrire que $\theta_{VM} \sim \mathrm{N}\big(\theta, \frac{\theta}{n}\big)$ (approximativement), car

$$E[X] = \mathrm{VAR}[X] = \theta$$

Remarque. On aurait aussi pu utiliser la remarque (ii), à la page 221, et l'exemple 5.2.5, à la page 220, pour obtenir ce résultat.

(b) Puisque la fonction $g(\theta) := 1/\theta = \beta$ possède la fonction inverse $g^{-1}(\beta) = 1/\beta$, on peut écrire que

$$\beta_{VM} = \frac{1}{\theta_{VM}} \overset{\text{(a)}}{=} \frac{1}{\bar{X}}$$

Question n° 23

Soit $X \sim \mathrm{B}(n, p)$. Trouver l'estimateur par la méthode des moments

(a) du paramètre p si $n = 10$;

(b) de n si $p = 1/2$.

Solution. (a) On pose:

$$E[X] = \bar{X} \iff 10\,p = \bar{X} \implies p_M = \frac{\bar{X}}{10}$$

(b) Dans ce cas,

$$E[X] = \bar{X} \iff \frac{n}{2} = \bar{X} \implies n_M = 2\bar{X}$$

Question n° 24

Soit $x_1, \ldots, x_{21}$ un échantillon aléatoire particulier de $X \sim \mathrm{N}\big(\mu, \sigma^2\big)$, où μ et σ^2 sont inconnus. Supposons que $\bar{x} = 1$ et $s = 0{,}2$. Calculer un intervalle de confiance (a) à 95 % pour μ; (b) à 99 % pour σ^2.

Solution. (a) On calcule

$$\bar{x} \pm t_{0{,}025;20} \frac{0{,}2}{\sqrt{21}} \overset{\text{tab. 5.2}}{\simeq} 1 \pm 2{,}086 \frac{1}{5\sqrt{21}} \simeq 1 \pm 0{,}091$$

(b) On calcule maintenant

$$\left[\frac{20s^2}{\chi^2_{0{,}005;20}}, \frac{20s^2}{\chi^2_{0{,}995;20}} \right] \overset{\text{tab. 5.3}}{\simeq} \left[\frac{20\,(0{,}04)}{40{,}00}, \frac{20\,(0{,}04)}{7{,}43} \right] \simeq [0{,}02; 0{,}108]$$

Question n° 25

Les données suivantes constituent un échantillon aléatoire particulier d'une variable aléatoire X désignant le nombre de défauts d'objets manufacturés: 2, 3, 1, 0, 4, 1, 2, 4, 1, 3.

(a) Calculer un intervalle de confiance à 95 % pour la proportion réelle θ des objets qui comptent au plus deux défauts.

(b) Combien d'observations, dans le pire des cas, devrait-on recueillir pour que la différence maximale, en valeur absolue, entre $\hat{\theta}$ et θ soit inférieure à 0,05, et ce, avec un coefficient de confiance de 95 %?

Solution. (a) On calcule

$$\bar{x} \pm z_{0,025}\sqrt{\frac{\bar{x}\,(1-\bar{x})}{10}} \simeq \frac{6}{10} \pm 1{,}960\sqrt{\frac{0{,}6 \times 0{,}4}{10}} \simeq 0{,}6 \pm 0{,}30$$

(b) On doit recueillir

$$n = \left(\frac{z_{0,025}}{0{,}05}\right)^2 \frac{1}{4} \simeq (20 \times 1{,}960)^2 \frac{1}{4} \simeq 384{,}2 \implies n_{\min} = 385$$

Question n$^{\text{o}}$ 26

Dans un sondage effectué auprès de 1000 personnes, 645 d'entre elles ont dit préférer l'option A aux autres options dans une question.

(a) Calculer un intervalle de confiance approximatif à 95 % pour la popularité p de l'option A.

(b) La population comprend deux strates, dont la plus grande fait 85 % du total. On peut alors écrire que

$$p = 0{,}15p_1 + 0{,}85p_2$$

où p_i est la popularité de l'option A dans la strate i, pour $i = 1, 2$, ce qui implique que

$$\hat{p} = 0{,}15\hat{p}_1 + 0{,}85\hat{p}_2$$

D'après le sondage, lequel est représentatif de la population, la popularité de l'option A est de 75 % dans la plus grande des deux strates. Calculer un intervalle de confiance à 95 % pour p qui tient compte de cette stratification.

Solution. (a) L'intervalle de confiance approximatif est donné par

$$\hat{p} \pm z_{0,025}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{1000}} \simeq 0{,}645 \pm 1{,}960\sqrt{\frac{0{,}645 \times 0{,}355}{1000}} \simeq 0{,}645 \pm 0{,}030$$

(b) On peut écrire que

$$\hat{p}_1 = \frac{1}{0{,}15}\left[0{,}645 - 0{,}85\,(0{,}75)\right] = 0{,}05$$

Alors on a:

$$\hat{p}_1 \approx \mathrm{N}\left(p_1, \frac{(0{,}05)(0{,}95)}{(0{,}15)(1000)}\right) \quad \text{et} \quad \hat{p}_2 \approx \mathrm{N}\left(p_2, \frac{(0{,}75)(0{,}25)}{(0{,}85)(1000)}\right)$$

De plus, on peut supposer que $\hat{p}_1$ et $\hat{p}_2$ sont des variables aléatoires indépendantes. Il s'ensuit que

$$\hat{p} \approx \mathrm{N}\left(p, \sigma_{\hat{p}}^2\right)$$

où

$$\sigma_{\hat{p}}^2 = (0{,}15)^2 \frac{0{,}05 \times 0{,}95}{150} + (0{,}85)^2 \frac{0{,}75 \times 0{,}25}{850} \simeq (0{,}013)^2$$

De là, on déduit qu'un intervalle de confiance approximatif à 95 % pour p est

$$0{,}645 \pm 1{,}960 \times 0{,}013 \simeq 0{,}645 \pm 0{,}0255$$

Remarque. Dans cet exercice, $\hat{p}$ désigne parfois l'estimateur du paramètre inconnu p, et parfois une estimation ponctuelle de p; de même pour $\hat{p}_i$. Il faut se rappeler qu'un estimateur est une variable aléatoire, tandis qu'une estimation ponctuelle est un nombre qui correspond au paramètre.

Question n$^{\text{o}}$ 27

Au cours d'une période d'environ six ans, on a dénombré, à l'échelle mondiale, 161 tremblements de terre d'une intensité de 4 ou plus sur l'échelle de Richter. On a calculé le nombre de jours écoulés entre deux tremblements de terre consécutifs. Les résultats sont les suivants:

Nombre de jours	0–4	5–9	10–14	15–19	20–24
Effectif	50	31	26	17	10

Nombre de jours	25–29	30–34	35–39	40 ou +
Effectif	8	6	6	7*

* Ces sept données sont: 40, 43, 44, 49, 58, 60 et 81 jours.

(a) Calculer approximativement la moyenne et l'écart-type de l'échantillon.

(b) On suppose que le nombre X de jours entre deux tremblements de terre consécutifs présente approximativement une distribution $\mathrm{N}(\mu, \sigma^2)$. Calculer un intervalle de confiance *unilatéral* approximatif (si nécessaire), avec borne supérieure,

(i) à 95 % pour μ;

(ii) à 99 % pour σ;

(iii) à 90 % pour $p := P[X \geq 25]$.

Indication. On a: $\chi^2_{0,99;160} \simeq 121{,}33$.

Solution. (a) On a:

$$\bar{x} \simeq \frac{1}{161}\left[50(2) + 31(7) + \ldots + 6(37) + 40 + 43 + \ldots + 81\right] = \frac{2143}{161} \simeq 13{,}31$$

et

$$\begin{aligned} s^2 &\simeq \frac{1}{161-1}\left[50(2-13{,}31)^2 + \ldots + 6(37-13{,}31)^2 + (40-13{,}31)^2\right. \\ &\qquad \left. + \ldots + (81-13{,}31)^2\right] \simeq 176{,}20 \\ \implies s &\simeq 13{,}27 \end{aligned}$$

(b) (i) L'intervalle de confiance unilatéral approximatif, avec borne supérieure, est donné par

$$\left(-\infty, \bar{x} + z_{0,05}\frac{s}{\sqrt{n}}\right] \simeq \left(-\infty; 13{,}31 + 1{,}645\frac{13{,}27}{\sqrt{161}}\right] \simeq (-\infty; 15{,}03]$$

(ii) L'intervalle (exact) recherché est

$$\left[0, \frac{\sqrt{n-1}s}{\sqrt{\chi^2_{0,99;n-1}}}\right] \simeq \left[0, \frac{\sqrt{160}(13{,}27)}{\sqrt{121{,}33}}\right] \simeq [0; 15{,}24]$$

Remarque. La valeur du centile $\chi^2_{0,99;160}$, selon l'approximation de Wilson-Hilferty, est d'environ 121,34, tandis que $\chi^2_{0,99;160} \simeq 120{,}66$ d'après l'approximation de Fisher (voir la page 240).

(iii) On calcule d'abord

$$\hat{p} = \frac{8+6+6+7}{161} \simeq 0{,}1677$$

Alors l'intervalle de confiance est donné par (environ)

$$\left[0, \hat{p} + z_{0,10}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right] \overset{\text{tab. 5.1}}{\simeq} \left[0; 0{,}1677 + 1{,}282\sqrt{\frac{0{,}1677(0{,}8323)}{161}}\right]$$
$$\simeq [0; 0{,}205]$$

Question n^{o} 28

La durée moyenne et l'écart-type de la durée des 50 dernières inondations qui sont survenues, dans une certaine région, sont de 4 et 5 jours, respectivement. Quelles sont l'espérance mathématique et la variance de la moyenne $\bar{X}$ d'un échantillon aléatoire *sans* remise de taille 10 tiré parmi ces 50 dernières inondations?

Solution. On a: $\mu_{\bar{X}} = \mu_X = 4$ et

$$\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}\left(\frac{N-n}{N-1}\right) = \frac{25}{10}\left(\frac{50-10}{50-1}\right) \simeq 2{,}04$$

Question n^{o} 29

Soit X_1 et X_2 deux variables aléatoires indépendantes. On suppose que X_i présente une distribution de Poisson de paramètres λ_i, pour $i = 1, 2$.

(a) On définit $\lambda = \lambda_1 - \lambda_2$. Pour estimer λ, on propose d'utiliser $\hat{\lambda} = X_1 - X_2$.

(i) L'estimateur $\hat{\lambda}$ est-il sans biais? Justifier.

(ii) Calculer la variance de $\hat{\lambda}$.

(b) Soit $X_{11}, \ldots, X_{1n}$ un échantillon aléatoire de taille n de la variable aléatoire X_1. Pour estimer le paramètre λ_1, on se sert de la moyenne $\bar{X}_1$ de l'échantillon aléatoire. Utiliser ce fait pour obtenir un intervalle de confiance (approximatif) à 95 % pour λ_1 si $n = 100$ et $\bar{x}_1 = 25$.

Solution. (a) (i) On a: $E\left[\hat{\lambda}\right] = E\left[X_1 - X_2\right] = \lambda_1 - \lambda_2 = \lambda$. Donc $\hat{\lambda}$ est sans biais.

(ii) On calcule

$$\mathrm{VAR}\left[\hat{\lambda}\right] = \mathrm{VAR}\left[X_1 - X_2\right] \overset{\text{ind.}}{=} \mathrm{VAR}\left[X_1\right] + \mathrm{VAR}\left[X_2\right] = \lambda_1 + \lambda_2$$

(b) On pose $\hat{\lambda}_1 = \bar{X}_1$. Or, par le théorème central limite, on peut écrire que

$$\bar{X}_1 \approx \mathrm{N}\left(\lambda_1, \frac{\lambda_1}{n}\right)$$

Il s'ensuit que

$$P\left[-1{,}960 \leq \frac{\bar{X}_1 - \lambda_1}{\sqrt{\lambda_1/n}} \leq 1{,}960\right] \simeq 0{,}95$$

$$\implies P\left[-1{,}960 \leq \frac{\bar{X}_1 - \lambda_1}{\sqrt{\hat{\lambda}_1/n}} \leq 1{,}960\right] \simeq 0{,}95 \quad (\text{car } \lambda_1 \text{ est inconnu})$$

$$\overset{\hat{\lambda}_1=\bar{X}_1}{\implies} P\left[\bar{X}_1 - 1{,}960\sqrt{\frac{\bar{X}_1}{n}} \leq \lambda_1 \leq 1{,}960\sqrt{\frac{\bar{X}_1}{n}}\right] \simeq 0{,}95$$

Avec $n = 100$ et $\bar{x}_1 = 25$, on obtient l'intervalle de confiance suivant: $[24{,}02; 25{,}98]$.

Question n$^{\text{o}}$ 30

Une machine fabrique des billes en métal que l'on classe (en fonction de leur diamètre) en trois catégories: A, B et C. Une bille est dite de catégorie A si son diamètre est inférieur à 9,5 mm, de catégorie B si son diamètre est au moins de 9,5 mm et au plus de 10,5 mm, et de catégorie C si son diamètre est supérieur à 10,5 mm.

Un échantillon aléatoire particulier de 200 billes est prélevé de la production de la machine et on y observe 12 billes de catégorie A, 156 de catégorie B et 32 de catégorie C. Construire un intervalle de confiance à 95 % pour la proportion réelle p_B de billes de catégorie B produites par la machine.

Solution. On a:

$$\hat{p}_B = \frac{156}{200}$$

Alors l'intervalle de confiance (approximatif) est donné par

$$\hat{p}_B \pm z_{0{,}025}\sqrt{\frac{\hat{p}_B\,(1-\hat{p}_B)}{200}} \simeq 0{,}780 \pm 0{,}057 \implies [0{,}723; 0{,}837]$$

Exercices

Question n$^{\text{o}}$ 1

Soit (x_i, y_i), $i = 1, 2, \ldots, n$, des observations particulières du couple de variables aléatoires (X, Y). On pose: $U = aX + cY$, où a et c sont deux constantes quelconques, et

$$u_i = ax_i + cy_i \quad \text{pour } i = 1, 2, \ldots, n$$

Soit $\bar{x}, \bar{y}, \bar{u}$ et s_x^2, s_y^2, s_u^2 les moyennes et les variances des observations de X, Y, U, et $s_{x,y}$ la covariance de observations de X et Y.

(a) Montrer que $\bar{u} = a\bar{x} + c\bar{y}$.

(b) Montrer que $s_u^2 = a^2 s_x^2 + 2acs_{x,y} + c^2 s_y^2$.

Question n^o 2

La quantité d'énergie utilisée par une certaine machine en une heure est une variable aléatoire qui présente une distribution (approximativement) normale de moyenne 50 et d'écart-type 1. Soit X_1 et X_2 la quantité d'énergie utilisée entre 10 h et 11 h, et entre 15 h et 16 h, respectivement, au cours d'une journée. On suppose que X_1 et X_2 sont indépendantes.

(a) Quelle est la distribution de probabilité de la variable $U := (X_1 - 50)^2$?

(b) Trouver un nombre réel u tel que $P[U \geq u] = 0{,}05$.

(c) Quelle est la distribution de probabilité de la variable

$$V := (X_2 - 50) / \sqrt{(X_1 - 50)^2}$$

(d) Trouver un nombre réel v tel que $P[V \leq v] = 0{,}01$.

Question n^o 3

Obtenir l'estimateur à vraisemblance maximale du paramètre inconnu θ (> 0) de la fonction de densité

$$f_X(x; \theta) = \theta^2 x e^{-\theta x} \text{ pour } x \geq 0$$

Question n^o 4

On a effectué une étude portant sur la durabilité des pneus de deux marques, A et B, en fonction de trois classes de prix: bas, moyen, élevé. Le tableau suivant est un classement du nombre de pneus selon la marque, le prix et la durée. La durée est mesurée en unités de 10.000 km.

	marque	A			B		
	durée	2	4	6	2	4	6
prix	bas	13	5	2	15	9	6
	moyen	5	12	3	10	10	10
	élevé	4	9	7	2	8	20

(a) Notons d_j la durée du pneu, où

$$d_j = \begin{cases} 2 \text{ si } j = 1 \\ 4 \text{ si } j = 2 \\ 6 \text{ si } j = 3 \end{cases}$$

Compléter le tableau suivant:

durée	2	4	6	total
marque				
A	$n_{A_1} =$	$n_{A_2} =$	$n_{A_3} =$	
B	$n_{B_1} =$	$n_{B_2} =$	$n_{B_3} =$	
total				

où n_{A_j} (respectivement n_{B_j}) est le nombre de pneus de marque A (respectivement B) ayant une durée d_j, pour $j = 1, 2, 3$.

(b) On calcule

$$\sum_{j=1}^{3} n_{A_j} d_j = 220, \quad \sum_{j=1}^{3} n_{A_j} d_j^2 = 936, \quad \sum_{j=1}^{3} n_{B_j} d_j = 378, \quad \sum_{j=1}^{3} n_{B_j} d_j^2 = 1836$$

Quelle marque possède

(i) la plus grande durée moyenne?

(ii) le plus petit coefficient de variation de durée?

Question n$^{\text{o}}$ 5

Soit $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ un échantillon aléatoire particulier, rangé en ordre croissant, d'une certaine variable aléatoire X. On pose:

$$y_i = \frac{x_{(i)} - \bar{x}}{s}$$

pour $i = 1, \ldots, n$, où $\bar{x}$ et s désignent la moyenne et l'écart-type de l'échantillon, respectivement.

(a) Montrer que $\bar{y} = 0$.

(b) Soit s_y l'écart-type de l'échantillon $y_1, y_2, \ldots, y_n$. Montrer que $s_y = 1$.

(c) Calculer l'écart-type de l'échantillon $x_1, x_2, \ldots, x_n$ (les *données brutes*) si

$$n = 5, \quad \sum_{i=1}^{5} x_{(i)} = 10, \quad \sum_{i=1}^{5} x_{(i)}^2 = 30$$

Question n° 6

Trouver la taille n d'un échantillon aléatoire provenant d'une population $\mathrm{N}(\mu, \sigma^2)$ de telle sorte que

$$P\left[\left|\frac{\bar{X}-\mu}{S/\sqrt{n}}\right| < 2{,}06\right] = 0{,}95$$

Question n° 7

Soit $x_1, x_2, \ldots, x_n$ une série d'observations particulières de la variable aléatoire X. On définit, pour $k = 1, 2, \ldots, n$,

$$\bar{x}_0 = 0 \quad \text{(par convention)}$$

$$\bar{x}_k = \frac{1}{k}\sum_{j=1}^{k} x_j \quad \text{(la moyenne des } k \text{ premières observations)}$$

$$d_k = x_k - \bar{x}_{k-1} \quad \text{(l'écart entre la } k^{\mathrm{e}} \text{ observation et la moyenne des } (k-1) \text{ premières observations)}$$

$$s_k^2 = \frac{1}{k-1}\sum_{j=1}^{k} (x_j - \bar{x}_k)^2 \quad \text{(la variance des } k \geq 2 \text{ premières observations)}$$

(a) Montrer que

(i) $\bar{x}_k = \bar{x}_{k-1} + \dfrac{d_k}{k}$;

(ii) $\sum_{j=1}^{k} (x_j - \bar{x}_k)^2 = \sum_{j=1}^{k-1} (x_j - \bar{x}_{k-1})^2 + \left(\frac{k-1}{k}\right) d_k^2$;

(iii) $s_k^2 = \left(\frac{k-2}{k-1}\right) s_{k-1}^2 + \dfrac{d_k^2}{k}$ pour $k \geq 2$.

(b) Utiliser les formules ci-dessus pour calculer la moyenne et la variance d'une série de 20 observations si

$$\bar{x}_{18} = 100, \quad s_{18}^2 = 900, \quad x_{19} = 109{,}5, \quad x_{20} = 101{,}5$$

Question n° 8

Un échantillon aléatoire particulier de taille $n = 5$ du couple de variables aléatoires (X, Y) a été recueilli. Calculer le coefficient de corrélation des données si l'on a obtenu les résultats suivants:

$$\sum_{i=1}^{5} x_i = -25, \quad \sum_{i=1}^{5} x_i^2 = 165, \quad \sum_{i=1}^{5} y_i = 30, \quad \sum_{i=1}^{5} y_i^2 = 230, \quad \sum_{i=1}^{5} x_i y_i = -194$$

Question n° 9

Soit X_1 et X_2 deux variables aléatoires $\mathrm{N}(0,1)$ indépendantes. Trouver la constante c telle que

$$P\left[0 \leq \frac{X_1}{|X_2|} \leq c\right] = 0{,}45$$

Question n° 10

Calculer le coefficient d'asymétrie des données suivantes:

$$x_1 = 2, \quad x_2 = 5, \quad x_3 = 8, \quad x_4 = 9, \quad x_5 = 1$$

Question n° 11

Soit X une variable aléatoire qui présente une distribution $\mathrm{N}(0,1)$. Trouver la valeur approximative du 20^{e} centile de X^2.

Question n° 12

On a recueilli 10 observations particulières d'une variable aléatoire X.

(a) Calculer l'écart-type des observations si $\sum_{i=1}^{10} x_i = 55$ et $\sum_{i=1}^{10} x_i^2 = 385$.

(b) Si l'on prélève une 11^{e} observation, quelle est la plus petite valeur que peut prendre l'écart-type des 11 observations? Justifier.

(c) Si $x_{(4)} = 4$ et $x_{(6)} = 6$, quelle est la plus petite et la plus grande valeur que peut prendre la médiane des 10 observations?

Question n° 13

Soit $X_1, X_2, \ldots, X_{50}$ des variables aléatoires indépendantes. On suppose que X_i présente une distribution du khi-deux à 10 degrés de liberté, pour $i = 1, 2, \ldots, 50$.

(a) Posons $Y = (X_1 + X_2)/2$. Trouver le nombre a tel que $P[Y \leq a] = 0{,}01$.

(b) Calculer approximativement $P\left[9{,}8 \leq \bar{X} < 10{,}1\right]$. Justifier l'approximation utilisée.

Question n° 14

Soit X_1, X_2 et X_3 des variables aléatoires indépendantes qui présentent toutes les trois une distribution $\mathrm{N}(0,1)$. On définit $U = \sqrt{X_2^2 + X_3^2}$. Trouver le nombre d tel que $P[X_1 > d\,U] = 0{,}10$.

Question n° 15

On dispose des observations particulières $x_1,\ x_2,\ \ldots, x_n$ d'une variable aléatoire X, où n est un nombre pair. On pose:

$$y_i = ax_i + b \quad \text{pour } i = 1, 2, \ldots, n,$$

où $a \neq 0$ et b sont des nombres réels.

(a) Soit $n = 2k$. Exprimer $y_{(k)}$ en fonction des x_i.

(b) Exprimer la médiane $\tilde{y}$ en fonction de $\tilde{x}$.

Question n° 16

La durée X (en heures) des pannes majeures d'électricité, dans une certaine région, présente une distribution (approximativement) normale de moyenne 4 et d'écart-type 2. On suppose que les pannes se produisent indépendamment les unes des autres. Soit $X_1, X_2, \ldots, X_{10}$ la durée des 10 prochaines pannes majeures. On pose:

$$\bar{X} = \frac{1}{5}\sum_{i=1}^{5} X_i, \quad U = \frac{1}{4}\sum_{i=1}^{5}\left(X_i - \bar{X}\right)^2, \quad V = \frac{1}{4}\sum_{i=6}^{10}(X_i - 4)^2$$

Trouver les nombres a, b et c tels que

(a) $P\,[U < a] = 0{,}10$.

(b) $P\,[U + V > b] = 0{,}01$.

(c) $P\,[U > c\,V] = 0{,}05$.

Question n° 17

Soit $X_1 \sim \mathrm{N}(0, 1)$ et $X_2 \sim \mathrm{N}(1, 3)$ deux variables aléatoires indépendantes.

(a) Trouver le nombre c tel que $P\left[(X_2 - 1)^2 < c\right] = 0{,}90$.

(b) Trouver le nombre d tel que

$$P\left[\frac{X_2 - 1}{\sqrt{X_1^2}} > \sqrt{3}d\right] = 0,01$$

Question n° 18

Soit X une variable aléatoire qui présente une distribution gamma de paramètres $\alpha = 25$ et $\lambda = 1/2$.

(a) Utiliser le fait que X présente aussi une distribution du khi-deux à 50 degrés de liberté pour calculer approximativement la probabilité $P\,[X < 30]$.

(b) Comparer la réponse en (a) avec la réponse exacte obtenue en se servant d'une distribution de Poisson.

Question n° 19

Soit X une variable aléatoire qui présente une distribution normale de paramètres $\mu = 50$ et $\sigma^2 = 25$. On définit $W = X - 50$. Trouver le nombre d tel que $P\left[W^2 > d\right] = 0{,}5$.

Question n° 20

Soit $X_1, \ldots, X_{400}$ un échantillon aléatoire de X, où X est une variable aléatoire qui présente une distribution de Student à 3 degrés de liberté. Utiliser le théorème central limite pour calculer approximativement $P\left[\bar{X} > 0{,}05\right]$.

Question n° 21

Supposons que $X \sim \mathrm{N}(0,1)$ et $Y \sim \chi^2_{10}$ sont deux variables aléatoires indépendantes. Trouver le nombre k tel que

$$P\left[\frac{10X^2}{Y} < k\right] = 0{,}05$$

Question n° 22

Soit $X_1, \ldots, X_n$ des variables aléatoires indépendantes qui présentent toutes une distribution $\mathrm{N}(0,1)$. Trouver les nombres n tels que

(a) $P\left[X_1 > 0, X_2 > 0, \ldots, X_n > 0\right] = 0{,}0625$;

(b) $P\left[\bar{X} < 0{,}329\right] \simeq 0{,}95$;

(c) $P\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 > 0{,}02\right] \simeq 0{,}99$.

Question n° 23

Supposons que les variables aléatoires $X_1, X_2, \ldots, X_{100}$ sont indépendantes et présentent toutes une distribution $\mathrm{N}(0,1)$.

(a) Calculer la probabilité $P\left[|X_1| > 0{,}018\,\left(X_2^2 + \ldots + X_{50}^2\right)^{1/2}\right]$.

(b) Utiliser le théorème central limite pour calculer approximativement la probabilité $P\left[X_1^2 + \ldots + X_{50}^2 > 2\,\left(X_{51}^2 + \ldots + X_{100}^2\right)\right]$.

Question n° 24

On a observé le nombre de voitures qui passent à une intersection au cours d'une période de cinq minutes. On a répété cette expérience à 20 reprises. Les données suivantes ont été recueillies:

17	17	27	19	18	19	19	16	16	20
28	21	23	15	30	23	17	26	15	27

(a) Calculer $\bar{x}$, s^2, $\hat{\beta}_1$ et $\hat{\beta}_2$ (à l'aide d'un logiciel, si possible).

(b) En se basant sur les données, quel est approximativement le 25^{e} centile de la distribution dont elles proviennent.

(c) Trouver une distribution, parmi celles que l'on a définies au chapitre 3, dont peuvent provenir les données. Justifier votre réponse.

Question n$^{\text{o}}$ 25

Soit X_1 et X_2 des variables aléatoires indépendantes qui présentent une distribution $\mathrm{N}(0,1)$ et une distribution $\mathrm{N}(0,2)$, respectivement. On prélève un échantillon aléatoire de taille $n_1 = 21$ de X_1 et un échantillon aléatoire de taille $n_2 = 81$ de X_2.

(a) Calculer $P\left[S_1^2 + S_2^2 < 3{,}5\right]$, où S_i^2 est la variance de l'échantillon aléatoire de X_i, pour $i = 1, 2$.

(b) Utiliser le théorème central limite pour calculer approximativement la probabilité $P[S_2^2 > 3]$. Justifier l'utilisation du théorème central limite.

Question n$^{\text{o}}$ 26

On s'intéresse au nombre de défauts que présentent des objets manufacturés. Des données recueillies à partir de 100 objets ont permis de constituer le tableau suivant:

Nombre de défauts	0	1	2	3	4	5
Nombre d'objets	20	25	45	9	0	1

(a) Calculer $\bar{x}$, s^2, $\hat{\beta}_1$ et $\hat{\beta}_2$.

(b) Soit θ la probabilité qu'un objet quelconque présente au plus deux défauts.

(i) Estimer la valeur de θ.

(ii) Trouver une valeur de θ_1 telle que $P[\theta_1 \leq \theta \leq 0{,}96] \simeq 0{,}95$.

(iii) Si l'on se base sur les 100 données recueillies, quel est approximativement le nombre de données dont on doit disposer pour pouvoir estimer θ avec une erreur maximale de 0,05, et ce, avec un coefficient de confiance de 95 %?

Question n$^{\text{o}}$ 27

On suppose que sur une certaine autoroute, entre 17 h et 19 h, le temps (en secondes) qui s'écoule entre le passage de deux voitures consécutives, allant dans le même direction, présente une distribution exponentielle dont la moyenne est

égale à 2. On suppose aussi que les temps d'attente entre les voitures sont des variables aléatoires indépendantes. Calculer la probabilité que

(a) le temps écoulé entre le passage de la première et de la deuxième voiture après 17 h soit inférieur au temps écoulé entre la deuxième et la troisième voiture;

(b) plus de deux minutes s'écoulent entre le passage de la première et de la 50^{e} voiture après 17 h

(i) de façon exacte;

(ii) en utilisant le théorème central limite;

(c) le temps écoulé entre le passage de la première et de la 20^{e} voiture après 17 h soit inférieur au temps écoulé entre le passage de la 20^{e} et de la 49^{e} voiture.

Remarque. Les questions (b) (i) et (c) nécessitent l'utilisation d'un logiciel.

Question n$^{\text{o}}$ 28

Vingt-cinq nombres aléatoires (indépendants) provenant d'une distribution $\mathrm{N}(0,1)$ sont générés par un ordinateur. Quelle est la probabilité que

(a) la valeur absolue de la somme des nombres obtenus soit supérieure à 5?

(b) la somme des carrés des nombres obtenus soit comprise entre 10 et 15?

(c) le résultat de la division du premier nombre par la racine carrée de la somme des carrés des 24 autres nombres soit supérieur à 1/2?

(d) le résultat de la division de la somme des carrés des 10 premiers nombres par la somme des carrés des 10 derniers nombres soit inférieur à 10?

Remarque. Les questions (b), (c) et (d) nécessitent l'utilisation d'un logiciel.

Question n$^{\text{o}}$ 29

Soit $x_1, \ldots, x_n$, n données. On définit:

$$y_i = \frac{x_i - \bar{x}}{s} \quad \text{pour } i = 1, \ldots, n$$

où $\bar{x}$ est la moyenne et s l'écart-type des n données. Calculer (a) le coefficient d'asymétrie et (b) le coefficient d'aplatissement des y_i en fonction de celui des x_i.

Question n$^{\text{o}}$ 30

On considère une population $\mathrm{N}(0,1)$. On y prélève 50 échantillons aléatoires indépendants de taille $n = 5$. Soit $S_1^2, \ldots, S_{50}^2$ les variances des échantillons. Calculer la probabilité $P\left[\sum_{i=1}^{50} S_i^2 > 40\right]$

(a) de façon exacte (à l'aide d'un logiciel);

(b) en utilisant le théorème central limite.

Question n° 31

Soit X une variable aléatoire qui présente une distribution $\mathrm{N}(0, \sigma^2)$, où σ est un paramètre inconnu. Pour estimer ce paramètre, on propose d'utiliser la statistique $\hat{\sigma} := |X_1|$, où X_1 est une observation de X.

(a) Quelle est l'erreur quadratique moyenne de $\hat{\sigma}$?

(b) Quelle est la probabilité que l'intervalle $[0{,}51\,|X_1|; 7{,}7\,|X_1|]$ contienne le paramètre σ?

Indication. Si $X \sim \mathrm{N}(0, \sigma^2)$, alors $E[|X|] = \sqrt{2/\pi}\,\sigma$.

Question n° 32

Soit

$$f_X(x;\theta) = \begin{cases} \theta e^{2\theta x} & \text{si } x < 0 \\ \theta e^{-2\theta x} & \text{si } x \geq 0 \end{cases}$$

où $\theta > 0$ est un paramètre inconnu. Estimer ce paramètre par la méthode de vraisemblance maximale.

Question n° 33

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de $X \sim \mathrm{N}(0, \sigma_X^2)$ et $Y_1, \ldots, Y_m$ un échantillon aléatoire de $Y \sim \mathrm{N}(0, \sigma_Y^2)$. On suppose que les deux échantillons sont indépendants. Obtenir une formule pour un intervalle de confiance à $100(1-\alpha)$ % pour le quotient σ_X^2/σ_Y^2.

Indication. Il faut utiliser le fait que μ_X et μ_Y sont connus ($\mu_X = \mu_Y = 0$).

Question n° 34

On veut estimer le paramètre p de la variable aléatoire X dont la fonction de probabilité est donnée par

$$p_X(x;p) = (1-p)^{x-1}p$$

où $0 < p < 1$ et $x = 1, 2, \ldots$ On dispose de deux observations indépendantes, X_1 et X_2, de la variable aléatoire X. On propose les deux estimateurs suivants:

$$\hat{p}_1 = \begin{cases} 1 \text{ si } X_1 = 1 \\ 0 \text{ autrement} \end{cases}$$

et

$$\hat{p}_2 = \begin{cases} 1 \text{ si } X_1 = 1 \text{ et } X_2 = 1 \\ \frac{1}{2} \text{ si } X_1 = 1 \text{ et } X_2 > 1, \text{ ou si } X_1 > 1 \text{ et } X_2 = 1 \\ 0 \text{ si } X_1 > 1 \text{ et } X_2 > 1 \end{cases}$$

(a) Montrer que les deux estimateurs sont non biaisés.

(b) Quel est l'estimateur (relativement) le plus efficace de p? Justifier votre réponse.

Question n° 35

On suppose que la distribution des revenus est une variable aléatoire X qui possède une fonction de densité $f_X(x;\theta)$ de la forme

$$f_X(x;\theta) = \begin{cases} 0 & \text{si } x < \theta \\ \dfrac{3\theta^3}{x^4} & \text{si } x \geq \theta > 0 \end{cases}$$

On montre que $E[X] = \frac{3\theta}{2}$ et $\text{VAR}[X] = \frac{3\theta^2}{4}$. Un estimateur de θ est fourni par $\hat{\theta} = \frac{2}{3}\bar{X}$.

(a) L'estimateur est-il sans biais? Justifier.

(b) Calculer son erreur quadratique moyenne.

(c) Utiliser le théorème central limite pour obtenir un intervalle de confiance approximatif à 95 % pour θ si $n = 100$ et $\bar{x} = 30$.

(d) Montrer que l'estimateur de θ obtenu par la méthode de vraisemblance maximale est

$$\theta_{VM} = \min\{X_1, \ldots, X_n\}$$

Question n° 36

Soit $X \sim \text{N}(0,1)$ et $Y \sim \text{N}(3,5)$ deux variables aléatoires indépendantes.

(a) Calculer $P[2X < Y]$.

(b) Trouver le 50e centile de $Z := X + Y$.

(c) Trouver le nombre c tel que $P\left[(Y-3)^2 < 5c\right] = 0{,}05$.

(d) Trouver le nombre d tel que $P\left[(Y-3)^2 > d\,X^2\right] = 0{,}01$.

Question n° 37

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n pris dans une population $\text{N}(\mu, \sigma^2)$, et soit

$$S = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]^{1/2} \quad \text{et} \quad S^* = \left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]^{1/2}$$

deux estimateurs de l'écart-type σ.

(a) Compléter le tableau suivant:

Estimateur	Biais	Variance
S		
S^*		

Indication. Si Y présente une distribution χ^2_{n-1}, alors on peut montrer que

$$E\left[\sqrt{Y}\right] = \sqrt{2}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

(b) Calculer l'erreur quadratique moyenne de S et celle de S^*.

(c) Lequel de S ou S^* est préférable comme estimateur de σ si l'on emploie le critère de l'erreur quadratique moyenne?

Question n$^{\text{o}}$ 38

On veut estimer le paramètre q de la fonction de probabilité suivante:

$$p_X(x;q) = (1-q)\,q^{x-1}$$

où $q > 0$ et $x = 1, 2, \ldots$, et ce, en utilisant une *seule* observation de la variable aléatoire X. Estimer q

(a) par la méthode du maximum de vraisemblance;

(b) par la méthode des moments.

Question n$^{\text{o}}$ 39

Une variable aléatoire continue X possède la fonction de densité

$$f_X(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta}\exp\left(-\frac{x-\alpha}{\beta}\right) & \text{si } x \geq \alpha \\ 0 & \text{ailleurs} \end{cases}$$

où $-\infty < \alpha < \infty$ et $\beta > 0$. On trouve que

$$E[X] = \alpha + \beta, \quad E\left[X^2\right] = (\alpha+\beta)^2 + \beta^2$$

et

$$B^2 := E\left[\frac{d}{d\beta}\ln f_X(x;\alpha,\beta)\right]^2 = \frac{1}{\beta^2}$$

(a) Supposons d'abord que α est connu et β inconnu.

(i) Obtenir β_{VM}, l'estimateur à vraisemblance maximale de β.

(ii) Calculer le biais de β_{VM}.

(iii) Calculer l'erreur quadratique moyenne de β_{VM}.

(iv) Évaluer approximativement $P\left[\beta_{VM} > 1{,}5\,\beta\right]$ si l'on utilise un échantillon aléatoire de taille 36.

(v) Combien d'observations doit-on recueillir, si l'on veut être certain à 99 % que la valeur absolue de l'erreur commise en estimant β par β_{VM} n'est pas supérieure à $\beta/2$?

(b) Supposons ensuite que α aussi est inconnu. Estimer α et β par la méthode des moments.

Question n$^{\text{o}}$ 40

Soit $X_1, X_2, \ldots, X_{n_1}$ et $Y_1, Y_2, \ldots, Y_{n_2}$ deux échantillons aléatoires indépendants provenant de populations normales $\mathrm{N}\big(\mu_X, \sigma^2\big)$ et $\mathrm{N}\big(\mu_Y, \sigma^2\big)$ dont les moyennes μ_X et μ_Y sont connues.

(a) Montrer que l'estimateur à vraisemblance maximale de σ^2, basé sur les deux échantillons à la fois, est

$$\hat{\sigma}^2 := \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1}\left(X_i - \mu_X\right)^2 + \sum_{i=1}^{n_2}\left(Y_i - \mu_Y\right)^2\right]$$

(b) Calculer le biais de $\hat{\sigma}^2$.

(c) Calculer l'erreur quadratique moyenne de $\hat{\sigma}^2$.

Question n$^{\text{o}}$ 41

Soit X une variable aléatoire discrète dont la fonction de probabilité est donnée par

$$p_X\left(x;\theta\right) = \begin{cases} 1/\theta \text{ si } x = 1, 2, \ldots, \theta & (\theta \in \mathbb{N}) \\ 0 \quad \text{autrement} \end{cases}$$

On a: $E\left[X\right] = \frac{\theta+1}{2}$ et $\mathrm{VAR}[X] = \frac{\theta^2-1}{12}$.

(a) Obtenir θ_M, l'estimateur de θ par la méthode des moments.

(b) Calculer l'erreur quadratique moyenne de θ_M.

(c) Obtenir, en utilisant le théorème central limite, un intervalle de confiance approximatif à 99 % pour θ.

(d) Combien d'observations doit-on recueillir, si l'on veut être certain à environ 95 % que l'erreur maximale commise en estimant θ par θ_M est égale à 2, si l'on croit que la vraie valeur de θ est 20?

(e) Obtenir l'estimateur à vraisemblance maximale de θ.

Question n° 42

Soit X une variable aléatoire discrète dont la fonction de probabilité est donnée par

$$p_X(x;\theta) = \binom{x-1}{r-1}\theta^{-r}\left(1-\frac{1}{\theta}\right)^{x-r} \quad \text{pour } x = r, r+1, \ldots$$

où r est un entier fixe connu et θ est un paramètre inconnu plus grand que 1. On montre que $E[X] = r\theta$ et $\text{VAR}[X] = r\theta(\theta-1)$.

(a) Trouver l'estimateur à vraisemblance maximale, θ_{VM}, de θ.

(b) Calculer l'erreur quadratique moyenne de θ_{VM}.

(c) Obtenir une formule pour un intervalle de confiance approximatif pour θ, avec un coefficient de confiance de $1-\alpha$, où $0 < \alpha < 1$.

(d) *Application numérique.* Calculer l'intervalle de confiance pour θ en (c) si $n = 100$, $r = 5$, $\alpha = 0{,}05$ et $\sum_{i=1}^{100} x_i = 600$.

Question n° 43

Soit X une variable aléatoire dont la fonction de densité est donnée par

$$f_X(x;\theta) = 2^{\theta}\theta x^{\theta-1} \quad \text{pour } 0 < x < \tfrac{1}{2}$$

où $\theta > 0$. Estimer le paramètre θ

(a) par la méthode de vraisemblance maximale;

(b) par la méthode des moments.

Question n° 44

La fonction de densité d'une variable aléatoire X qui présente une distribution exponentielle est donnée par

$$f_X(x;\theta) = \theta e^{-\theta x} \quad \text{pour } x > 0$$

où $\theta > 0$. Sachant que l'estimateur à vraisemblance maximale du paramètre θ est

$$\theta_{VM} = 1/\bar{X}$$

obtenir un intervalle de confiance approximatif à 99 % pour θ, basé sur un échantillon aléatoire de 36 observations.

Question n^{o} 45

Soit $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, où μ est un paramètre *connu*. Pour estimer le paramètre inconnu σ^2, on propose deux estimateurs:

$$\hat{\sigma}_1^2 := \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \mu\right)^2 \text{ et } \hat{\sigma}_2^2 := \frac{1}{n+1} \sum_{i=1}^{n} \left(X_i - \mu\right)^2$$

Quel estimateur possède la plus petite erreur quadratique moyenne?

Question n^{o} 46

Soit

$$f_X\left(x; \theta\right) = \begin{cases} \frac{1}{2}\theta e^{\theta x} & \text{si } x < 0 \\ \frac{1}{2}\theta e^{-\theta x} & \text{si } x \geq 0 \end{cases}$$

où $\theta > 0$. On peut montrer que $E\left[X\right] = 0$ et $\mathrm{VAR}[X] = \frac{2}{\theta^2}$. Estimer θ

(a) par la méthode de vraisemblance maximale;

(b) par la méthode des moments.

Question n^{o} 47

Supposons que X est une variable aléatoire qui présente une distribution uniforme sur l'intervalle $(0, \theta)$; c'est-à-dire que

$$f_X\left(x, \theta\right) = \begin{cases} \frac{1}{\theta} \text{ si } 0 < x < \theta \\ 0 \text{ ailleurs} \end{cases}$$

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n de X. L'estimateur à vraisemblance maximale du paramètre inconnu θ est donné par

$$\theta_{VM} = X_{(n)} := \max\{X_1, \ldots, X_n\}$$

On peut montrer que la fonction de densité de $Y := X_{(n)}/\theta$ est

$$f_Y\left(y\right) = \begin{cases} n y^{n-1} \text{ si } 0 < y < 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

(a) Calculer le biais et l'erreur quadratique moyenne de θ_{VM}.

(b) Utiliser la fonction de répartition de Y pour obtenir un intervalle de confiance (exact) à 95 % pour θ, basé sur un échantillon aléatoire de taille $n = 10$.

Question n° 48

Soit X une variable aléatoire dont la fonction de densité est

$$f_X(x;\theta) = \begin{cases} 0 & \text{si } x \leq 0 \\ \dfrac{3x^2}{\theta} e^{-x^3/\theta} & \text{si } x > 0 \end{cases}$$

où θ est un paramètre positif. On peut montrer que $\frac{2}{\theta}X^3 \sim \chi^2_2$.

(a) Calculer θ_{VM}, l'estimateur à vraisemblance maximale de θ.

(b) Calculer l'erreur quadratique moyenne de l'estimateur obtenu en (a).

(c) Établir une formule donnant un intervalle de confiance *exact* pour θ avec un coefficient de confiance de $100\,(1-\alpha)$ %.

Indication. On a: $\frac{2n}{\theta}\theta_{VM} \sim \chi^2_{2n}$.

(d) L'estimateur θ_{VM} présente, pour n grand, une distribution approximativement normale. Trouver alors n tel que

$$P\left[|\theta_{VM} - \theta| < \frac{\theta}{10}\right] \geq 0{,}95$$

Question n° 49

Un sac contient un nombre indéterminé, $N+1$, de jetons numérotés de 0 à N.

(a) Estimer N par la méthode de vraisemblance maximale.

(b) Estimer N par la méthode des moments et calculer l'erreur quadratique moyenne de l'estimateur.

Indication. Soit X le numéro d'un jeton tiré au hasard parmi les $N+1$ jetons dans le sac; alors on a:

$$p_X(x) = \frac{1}{N+1} \quad \text{pour } x = 0, 1, \ldots, N$$

(c) On tire, au hasard et avec remise, cinq jetons et on obtient les nombres suivants: 15, 3, 5, 8 et 3. Entre les estimateurs calculés en (a) et (b), quel est le meilleur estimateur de N? Justifier votre réponse.

Question n° 50

Un générateur de nombres aléatoires est censé produire des entiers de 0 à 9 de façon que chaque chiffre ait la même probabilité d'apparaître. Les 10.000 premiers chiffres générés ont permis de constituer le tableau d'effectifs suivant:

0	1	2	3	4	5	6	7	8	9
967	1008	975	1022	1003	989	1001	981	1043	1011

(a) Obtenir un intervalle de confiance à 95 % pour la probabilité qu'un chiffre généré soit inférieur à 3.

(b) Quelle est la médiane des 10.000 chiffres générés?

Question n° 51

Une variable aléatoire X possède une fonction de densité $f_X(x;\theta)$, où θ est un paramètre inconnu. Un échantillon aléatoire de X est prélevé. Deux statistiques, $T_1(X_1,\ldots,X_n)$ et $T_2(X_1,\ldots,X_n)$, sont disponibles pour estimer θ. La moyenne et la variance de ces statistiques sont données dans le tableau suivant:

	Moyenne	Variance
$T_1(X_1,\ldots,X_n)$	θ	Δ_1
$T_2(X_1,\ldots,X_n)$	$\theta+\epsilon$	Δ_2

En fonction des quantités ϵ, Δ_1 et Δ_2, discuter du choix de la meilleure statistique pour estimer θ.

Question n° 52

Une variable aléatoire X possède la fonction de densité

$$f_X(x;\theta) = (\theta+1)(2-x)^{\theta} \quad \text{si } 1 < x < 2$$

où $\theta > -1$ est un paramètre inconnu. Étant donné un échantillon aléatoire $X_1,\ldots,X_n$ de X, trouver l'estimateur à vraisemblance maximale de θ.

Question n° 53

Le fabricant d'un nouvel engrais chimique prétend que, si l'on utilise son engrais, environ 80 % des graines qui ont été semées vont germer. Soit θ la proportion véritable des graines qui germent à l'aide de cet engrais. Dans une expérience réalisée avec 50 graines, on a trouvé que 35 d'entre elles avaient germé.

(a) Estimer θ par la méthode des moments.

(b) Obtenir un intervalle de confiance à 95 % pour θ.

Questions à choix multiple

Question n$^{\text{o}}$ 1

Calculer le coefficient de variation des 10 observations particulières, x_1, x_2, ..., x_{10}, d'une variable aléatoire X, si l'on a trouvé que $\sum_{i=1}^{10} x_i = 30$ et $\sum_{i=1}^{10} x_i^2 = 110$.

(a) 0,50 (b) 0,74 (c) 1 (d) 1,50 (e) 2,22

Question n$^{\text{o}}$ 2

On dispose du tableau d'effectifs suivant:

Intervalle	$[0, 100)$	$[100, 200)$	$[200, 300)$	$[300, 400)$	$[400, 500)$	$[500, \infty)$
Effectif	5	20	25	25	20	5

Que peut-on affirmer au sujet de la moyenne $\bar{x}$ et de la médiane $\tilde{x}$ des données?

(a) $\bar{x} < \tilde{x}$ (b) $\bar{x} > \tilde{x}$ c) $\bar{x} \simeq \tilde{x}$ (d) $\bar{x} \neq \tilde{x}$ (e) rien

Question n$^{\text{o}}$ 3

On a recueilli un échantillon aléatoire particulier de taille 10, $x_1, \ldots, x_{10}$, d'une variable aléatoire X. On a calculé $\bar{x} = 0$ et $s = 1$. On prélève une 11$^{\text{e}}$ observation, x_{11}. Quelle est la valeur de x_{11}^2 si l'écart-type des 11 observations est aussi de 1?

(a) 0 (b) 10/11 (c) 1 (d) 11/10 (e) 10/9

Question n$^{\text{o}}$ 4

La moyenne et l'écart-type des données $x_1, \ldots, x_n$ sont $\bar{x} = 5$ et $s = 2$. Calculer la moyenne des carrés de ces données.

(a) $\dfrac{25}{n}$ (b) $25 - \dfrac{4}{n}$ (c) 25 (d) $29 - \dfrac{4}{n}$ (e) 29

Question n$^{\text{o}}$ 5

On nous donne les 10 observations particulières suivantes d'une variable aléatoire X: $61, 63, 58, 69, 72, 62, 59, 52, 52, 59$. Calculer la médiane des observations.

(a) 59 (b) 60 (c) 61 (d) 62 (e) 67

Question n$^{\text{o}}$ 6

On a recueilli 100 observations particulières d'une variable aléatoire X qui peut prendre les valeurs 1, 2, 3, 4, 5 et 6. À partir de ces observations, on a construit le tableau d'effectifs suivant:

Valeur	1	2	3	4	5	6
Effectif	15	20	12	25	13	15

Ensuite, 50 autres observations particulières de X ont été obtenues. Le coefficient de variation et l'écart-type de ces 50 nouvelles observations sont CV $= 0{,}5$ et $s = 1{,}745$. Calculer la moyenne de toutes les (150) observations.

(a) 3,46 (b) 3,47 (c) 3,48 (d) 3,49 (e) 3,50

Question n° 7

Soit X_1, X_2, X_3 un échantillon aléatoire de $X \sim \mathrm{N}(0,1)$ et Y_1, Y_2, Y_3, Y_4 un échantillon aléatoire de $Y \sim \mathrm{N}(0,4)$, où X et Y sont des variables aléatoires indépendantes. On pose: $W = \sum_{i=1}^{3} \left(X_i - \bar{X}\right)^2$.

(A) Calculer $P\left[\frac{\bar{Y}^2}{\bar{X}^2} < 484{,}2\right]$.

(a) 0,01 (b) 0,05 (c) 0,90 (d) 0,95 (e) 0,99

(B) Quelle distribution présente la variable aléatoire $W + \bar{Y}^2$?

(a) χ^2_1 (b) χ^2_2 (c) χ^2_3 (d) χ^2_4 (e) aucune distribution connue

(C) Laquelle (ou lesquelles) des statistiques suivantes présente(nt) une distribution de Student?

$$\text{I) } \sqrt{6}\frac{\bar{X}}{\sqrt{W}}, \quad \text{II) } \sqrt{2}\frac{\bar{Y}}{\sqrt{W}}, \quad \text{III) } \sqrt{3}\frac{\bar{X}}{\sqrt{\bar{Y}^2}}$$

(a) I seulement (b) II seulement (c) III seulement (d) I et II seulement
(e) I, II et III

Question n° 8

Soit

$$1{,}1 \quad 1{,}3 \quad 1{,}9 \quad 2{,}1 \quad 2{,}5 \quad 3{,}4$$

un échantillon aléatoire particulier de $X \sim \mathrm{N}(\mu, \sigma^2)$.

(A) Calculer un intervalle de confiance à 95 % pour μ.

(a) $2{,}05 \pm 1{,}645\frac{0{,}8385}{\sqrt{6}}$ (b) $2{,}05 \pm 1{,}960\frac{0{,}7030}{\sqrt{6}}$ (c) $2{,}05 \pm 1{,}960\frac{0{,}8385}{\sqrt{6}}$
(d) $2{,}05 \pm 2{,}015\frac{0{,}8385}{\sqrt{6}}$ (e) $2{,}05 \pm 2{,}571\frac{0{,}8385}{\sqrt{6}}$

(B) Calculer la limite supérieure d'un intervalle de confiance unilatéral à 99 % avec borne supérieure pour σ^2.

(a) $\frac{5\,(0{,}7030)}{16{,}75}$ (b) $\frac{5\,(0{,}7030)}{15{,}09}$ (c) $\frac{5\,(0{,}8385)}{15{,}09}$ (d) $\frac{5\,(0{,}7030)}{0{,}55}$ (e) $\frac{5\,(0{,}8385)}{0{,}55}$

(C) Calculer un intervalle de confiance à 95 % pour la probabilité $P[X > 2]$.

(a) $\frac{1}{2} \pm 1{,}645\left(\frac{1}{2}\right)$ (b) $\frac{1}{2} \pm 1{,}960\left(\frac{1}{2}\right)$ (c) $\frac{1}{2} \pm 1{,}645\left(\frac{1}{24}\right)^{1/2}$

(d) $\frac{1}{2} \pm 1{,}960\left(\frac{1}{24}\right)^{1/2}$ (e) $\frac{1}{2} \pm 2{,}571\left(\frac{1}{24}\right)^{1/2}$

Question n° 9

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de la variable aléatoire X dont la fonction de densité est

$$f_X(x;\theta) = \frac{2x}{\theta^2} \quad \text{si } 0 \leq x \leq \theta \quad (= 0 \text{ ailleurs})$$

On a: $\text{VAR}[X] = \frac{\theta^2}{18}$.

(A) On propose l'estimateur suivant du paramètre inconnu θ: $\hat{\theta} = \bar{X}$. Calculer l'erreur quadratique moyenne de $\hat{\theta}$.

(a) $\frac{\theta^2}{18n} - \frac{\theta}{3}$ (b) $\frac{\theta^2}{18n}$ (c) $\frac{\theta}{3} + \frac{\theta^2}{18n}$ (d) $\frac{\theta^2}{9n} + \frac{\theta^2}{18n}$ (e) $\frac{\theta^2}{9} + \frac{\theta^2}{18n}$

(B) Calculer l'estimateur à vraisemblance maximale de θ.

(a) $\bar{X}$ (b) $\frac{2}{3}\bar{X}$ (c) $\frac{3}{2}\bar{X}$ (d) $\min\{X_1, \ldots, X_n\}$ (e) $\max\{X_1, \ldots, X_n\}$

(C) Calculer l'estimateur de θ par la méthode des moments.

(a) $\bar{X}$ (b) $\frac{2}{3}\bar{X}$ (c) $\frac{3}{2}\bar{X}$ (d) $\min\{X_1, \ldots, X_n\}$ (e) $\max\{X_1, \ldots, X_n\}$

Question n° 10

On considère un échantillon aléatoire, $X_1, \ldots, X_n$, d'une variable aléatoire X dont la fonction de densité est donnée par

$$f_X(x;\theta) = \begin{cases} \theta e^{-\theta(x-1)} & \text{si } x \geq 1 \\ 0 & \text{ailleurs} \end{cases}$$

où $\theta > 0$ est un paramètre inconnu. On peut montrer que $E[X] = \frac{1}{\theta} + 1$ et $\text{VAR}[X] = \frac{1}{\theta^2}$.

(A) Pour estimer le paramètre θ, on propose d'utiliser $\hat{\theta} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculer le biais de $\hat{\theta}$.

(a) 0 (b) $1 - \theta$ (c) 1 (d) $\frac{1}{\theta} + 1 - \theta$ (e) $\frac{\theta + 1}{n} - \theta$

(B) Calculer l'erreur quadratique moyenne de $\hat{\theta}$ si $\theta = 1$.

(a) $\left(\frac{1}{n} - 1\right)^2$ (b) $\frac{1}{n} + 1$ (c) 1 (d) $\frac{1}{n^2} + 1$ (e) $\frac{1}{n}$

(C) Obtenir l'estimateur à vraisemblance maximale de θ.

(a) $\bar{X} - 1$ (b) $\sum_{i=1}^{n} X_i - n$ (c) $\dfrac{1}{\bar{X} - 1}$ (d) $\dfrac{1}{\bar{X}}$ (e) $\dfrac{1}{\sum_{i=1}^{n} X_i - n}$

Question n^{o} 11

On suppose qu'une variable aléatoire X présente une distribution de Poisson de paramètre λ. Le tableau suivant résume les observations disponibles de la variable X:

Valeur	0	1	2	3	4^{+}
Effectif	260	150	70	20	0

(A) Estimer la valeur du paramètre λ par la méthode des moments.

(a) 0,3 (b) 0,5 (c) 0,7 (d) 1 (e) 1,5

(B) Calculer un intervalle de confiance à 95 % pour la probabilité $P[X \leq 1]$.

(a) $0{,}18 \pm 1{,}645\sqrt{\frac{0{,}18 \times 0{,}82}{500}}$ (b) $0{,}18 \pm 1{,}960\sqrt{\frac{0{,}18 \times 0{,}82}{500}}$

(c) $0{,}82 \pm 1{,}645\sqrt{\frac{0{,}82 \times 0{,}18}{500}}$ (d) $0{,}30 \pm 1{,}960\sqrt{\frac{0{,}30 \times 0{,}70}{500}}$

(e) $0{,}82 \pm 1{,}960\sqrt{\frac{0{,}82 \times 0{,}18}{500}}$

Question n^{o} 12

Supposons que X_1, X_2 et X_3 sont trois variables aléatoires indépendantes telles que $X_1 \sim \mathrm{N}(0, 9)$, $X_2 \sim \mathrm{N}(-2, 4)$ et $X_3 \sim \mathrm{N}(0, 1)$. On définit

$$Y_1 = \frac{X_1}{3} \quad \text{et} \quad Y_2 = \frac{X_2 + 2}{2}$$

(A) Calculer $P\left[\frac{\sqrt{2}X_3}{\sqrt{Y_1^2 + Y_2^2}} > 1{,}886\right]$.

(a) 0,01 (b) 0,025 (c) 0,05 (d) 0,10 (e) 0,90

(B) Calculer $P\left[\sqrt{Y_1^2 + Y_2^2 + X_3^2} \leq 0{,}469\right]$.

(a) 0,025 (b) 0,05 (c) 0,10 (d) 0,90 (e) 0,975

6

Tests d'hypothèses

La théorie des tests d'hypothèses est l'un des sujets les plus importants en statistique mathématique. D'abord, les tests *d'ajustement* permettent d'affirmer si les données que l'on a recueillies semblent provenir d'une distribution donnée, par exemple d'une distribution normale. Ensuite, si la distribution normale est effectivement un modèle qui semble réaliste pour les données, on peut vouloir tester des hypothèses au sujet des paramètres de cette distribution. Nous allons aussi voir, en particulier, comment tester si la probabilité de succès dans une série d'essais de Bernoulli est égale à une constante donnée ou non, et si deux variables sont indépendantes.

Notons que l'on ne peut pas *prouver* qu'une hypothèse particulière est vraie en se servant d'un test d'hypothèses. Il y a toujours un risque de prendre la mauvaise décision en se basant sur les données. Par contre, lorsque le risque de commettre une erreur est inférieur à 1 % en rejetant une certaine hypothèse, par exemple, on peut raisonnablement conclure que cette hypothèse est fausse. De toute façon, lorsqu'on cherche un modèle pour des données, il est en pratique impossible de trouver le modèle *exact* si celui-ci est une distribution continue. Ainsi, en recueillant suffisamment d'observations d'une variable aléatoire continue, on peut arriver à conclure que n'importe quelle distribution que l'on propose pour cette variable n'est pas la bonne. De même, il est impossible que la moyenne d'une distribution normale quelconque soit *exactement* égale à une constante donnée, car la moyenne est un paramètre qui peut théoriquement prendre n'importe quelle valeur réelle.

6.1 Introduction et terminologie

Soit $x_1, x_2, \ldots, x_n$ un échantillon aléatoire particulier d'une variable aléatoire X qui possède une fonction de densité (ou une fonction de probabilité) f_X. On a deux **hypothèses** concernant la fonction f_X. Ces hypothèses peuvent être au sujet de f_X elle-même ou encore à propos d'un paramètre qui apparaît dans la fonction f_X. Après avoir calculé une statistique T, c'est-à-dire une fonction des données, on **accepte** une hypothèse et on **rejette** l'autre.

Pour tester l'hypothèse **nulle** H_0 contre la **contre-hypothèse** H_1 (aussi appelée hypothèse **alternative** par certains auteurs), on définit un **test** à l'aide d'une **région de rejet** C: ainsi, si $T(x_1, \ldots, x_n) \in C$, alors on rejette H_0 (et on accepte H_1), tandis que si $T(x_1, \ldots, x_n) \notin C$, alors on ne rejette pas (donc, on accepte) H_0. Le complément de C est appelé **région d'acceptation** du test.

Soit θ un paramètre inconnu qui apparaît dans la fonction f_X. Une hypothèse de la forme

$$H_0: \theta = \theta_0 \tag{6.1}$$

est une hypothèse **simple**, tandis que

$$H_1: \theta \neq \theta_0 \tag{6.2}$$

est une hypothèse **multiple** ou **composite**. Dans le cadre de ce livre (sauf dans l'exemple suivant), l'hypothèse H_0 sera toujours une hypothèse simple, tandis que H_1 sera toujours une hypothèse multiple.

Exemple 6.1.1.
(i) Supposons qu'on possède une pièce de monnaie truquée, pour laquelle $P[\{\text{face}\}] = 0{,}7$ *ou bien* $P[\{\text{pile}\}] = 0{,}7$. Dans ce cas, on a deux hypothèses alternatives simples:

$$H_0: p = 0{,}7 \quad \text{et} \quad H_1: p = 0{,}3$$

où p est la probabilité de l'événement $\{\text{face}\}$. En lançant la pièce de monnaie un assez grand nombre de fois, nous pourrons décider quelle hypothèse est vraie (ou, en fait, laquelle semble vraie).

(ii) Si l'on croit qu'une pièce de monnaie est truquée, alors on peut tester les hypothèses suivantes:

$$H_0: p = \tfrac{1}{2} \quad \text{contre} \quad H_1: p \neq \tfrac{1}{2}$$

(iii) Supposons qu'une certaine machine produit, en moyenne, c articles par jour et que la production journalière X d'une nouvelle machine, qui pourrait remplacer la première, est une variable aléatoire qui présente (approximativement) une distribution $\mathrm{N}(\mu, \sigma^2)$. On peut vouloir tester

$$H_0\colon \mu \leq c \quad \text{contre} \quad H_1\colon \mu > c$$

♢

Remarque. Dans le cas des tests *au sujet des paramètres* d'une variable aléatoire, l'hypothèse nulle est généralement celle que l'on croit *fausse* (particulièrement dans le cas d'un test *unilatéral*, comme ci-dessus). Par conséquent, on essaie de la rejeter. C'est pourquoi on préfère souvent dire que l'on ne peut pas rejeter H_0, plutôt que de dire que l'on accepte H_0.

6.1.1 Les espèces d'erreurs

Il y a deux espèces d'erreurs que l'on peut commettre:
(a) **erreur de première espèce**: accepter H_1 lorsque H_0 est vraie; c'est-à-dire rejet erroné de H_0;
(b) **erreur de seconde espèce**: accepter H_0 lorsque H_1 est vraie; c'est-à-dire acceptation erronée de H_0.

On définit les paramètres α et β comme suit:

$$\alpha = P[\text{Erreur de première espèce}] = P[\text{Rejeter } H_0 \mid H_0 \text{ est vraie}] \tag{6.3}$$

et

$$\beta = P[\text{Erreur de deuxième espèce}] = P[\text{Accepter } H_0 \mid H_0 \text{ est fausse}] \tag{6.4}$$

On appelle α **seuil** (**de signification**) ou **niveau** du test. De plus, la quantité $1-\beta$ est appelée **puissance** du test. Finalement, le graphique de $\beta(\theta)$ en fonction de θ est appelé **courbe** ou **abaque caractéristique** du test de $H_0\colon \theta = \theta_0$ contre $H_1\colon \theta \neq \theta_0$ (ou $H_1\colon \theta > \theta_0$, etc.). Notons que $\beta(\theta_0) = 1 - \alpha$.

Remarques.
(i) La quantité β dépend de la taille n de l'échantillon aléatoire: plus n augmente et plus β diminue. De plus, contrairement à α, la valeur de β n'est pas unique dans un problème donné mais dépend de l'hypothèse particulière H_1 que l'on suppose vraie. Il y a en fait une infinité d'hypothèses particulières H_1, tandis que H_0 est unique (dans le cas où H_0 est une hypothèse *simple*).

(ii) La notation des probabilités conditionnelles du chapitre 2 a été utilisée dans la définition des paramètres α et β; cependant, nous n'aurons jamais à calculer la probabilité de l'événement $\{H_0 \text{ est vraie}\}$. Il s'agit simplement d'une proposition que l'on suppose vraie dans le calcul du *risque de première espèce*. Il en est de même pour β.

(iii) Avec le choix que nous avons fait au sujet de la formulation des hypothèses, commettre une erreur de première espèce est généralement plus grave que commettre une erreur de seconde espèce. Ainsi, dans l'exemple 6.1.1 (iii), si l'on rejette H_0 de façon erronée, alors on recommandera d'acheter une nouvelle machine, laquelle aura une production inférieure (en moyenne) à celle de l'ancienne. Dans le cas contraire, on peut affirmer qu'il n'y a pas suffisamment d'indication (ou de preuve) statistique pour conclure que la nouvelle machine aura une production supérieure, et l'on préfère s'en tenir au *statu quo*. Une erreur a certes été commise, mais elle est moins coûteuse que la précédente. En pratique, il faudrait aussi tenir compte du coût d'achat de la nouvelle machine pour pouvoir décider si son acquisition est justifiée ou non.

Un autre exemple qui illustre bien l'importance de ne pas commettre une erreur de première espèce est celui d'un procès pour meurtre. Supposons que vous faites partie du jury qui doit décider si une personne accusée de meurtre est coupable ou non. Supposons aussi que vous croyez que la personne accusée est en fait coupable, et posons:

$$H_0\text{: la personne est innocente} \quad \text{et} \quad H_1\text{: la personne est coupable}$$

Dans ce cas, commettre une erreur de première espèce signifie condamner une personne innocente, tandis que commettre une erreur de seconde espèce se traduit par innocenter une personne qui était en fait coupable de meurtre. La plupart des gens reconnaîtront qu'il n'y a rien de pire que de condamner un innocent (surtout si la peine de mort est en vigueur...). Notons que, dans cet exemple, si l'on croit que la personne accusée est innocente, alors on pourrait poser:

$$H_0\text{: la personne est coupable} \quad \text{et} \quad H_1\text{: la personne est innocente}$$

mais choisir une valeur de β presque égale à 0.

(iii) Compte tenu de la remarque précédente, on choisit généralement le paramètre α dans l'intervalle $[0{,}01; 0{,}10]$. Ensuite, si l'échantillon aléatoire n'a pas encore été prélevé, on peut fixer une valeur du paramètre β; pour ce faire, il faut choisir un cas particulier de l'hypothèse H_1 pour lequel on voudrait que,

par exemple, il n'y ait pas plus de 15 % de risque d'accepter H_0 dans ce cas. La valeur de β que l'on fixe déterminera la taille de l'échantillon à prélever.

Par exemple, on pourrait poser que $\beta(\mu = 1{,}5c) = 0{,}10$ dans l'exemple 6.1.1 (iii); c'est-à-dire que, si la production moyenne de la nouvelle machine est supérieure de 50 % à celle de l'ancienne machine, alors on voudrait que le test utilisé n'ait pas plus de 10 % de risque de conclure que l'ancienne machine est au moins aussi bonne que la nouvelle.

6.2 Tests d'ajustement

6.2.1 Test d'ajustement du khi-deux de Pearson

Soit X une variable aléatoire dont la fonction de densité (ou la fonction de probabilité) $f_X(x;\theta)$ est inconnue. On veut faire le test de l'hypothèse nulle

$$H_0\colon f_X = f_0 \tag{6.5}$$

contre la contre-hypothèse

$$H_1\colon f_X \neq f_0 \tag{6.6}$$

où f_0 est une fonction donnée. La *marche à suivre* est la suivante:

(i) on divise l'ensemble des valeurs possibles de X en k *classes* (ou intervalles) disjointes et exhaustives;

(ii) on prélève un échantillon aléatoire de taille n de X;

(iii) on calcule

$$D^2 := \sum_{j=1}^{k} \frac{(n_j - m_j)^2}{m_j} \tag{6.7}$$

où n_j est le nombre d'observations dans la j^{e} classe (l'*effectif observé*) et m_j désigne l'*effectif sous* H_0, c'est-à-dire le nombre d'observations que devrait compter, *en moyenne*, la j^{e} classe si l'hypothèse H_0 est vraie;

(iv) on peut montrer que si H_0 est vraie, alors $D^2 \approx \chi^2_{k-r-1}$; c'est-à-dire que D^2 présente (approximativement) une distribution du khi-deux à $k - r - 1$ degrés de liberté, où r est le nombre de paramètres inconnus de la fonction f_0 qu'il *faut* estimer. Le test consiste à rejeter H_0 au seuil de signification (ou niveau) α si et seulement si

$$D^2 > \chi^2_{\alpha,k-r-1} \tag{6.8}$$

où (voir le chapitre 5, page 239)

$$P[X \leq \chi^2_{\alpha,n}] = 1 - \alpha \quad \text{si } X \sim \chi^2_n \tag{6.9}$$

Remarques.
(i) En général, il est préférable que m_j soit au moins égal à 5 pour tout j (certains auteurs donnent plutôt le critère $m_j \geq 3\ \forall j$, ou un autre critère semblable); si ce n'est pas le cas, alors on peut regrouper des classes et réduire par conséquent k.

(ii) Les k intervalles (ou classes) ne sont pas nécessairement de la même longueur; de même, il n'est pas nécessaire que les probabilités correspondantes soient égales.

(iii) Avant de prélever l'échantillon aléatoire, on peut affirmer que si l'hypothèse nulle H_0 est vraie, alors le nombre N_j d'observations qui seront situées dans la j^{e} classe est une variable aléatoire qui présente une distribution binomiale de paramètres n et p_j, où p_j est la probabilité que la variable X prenne une valeur dans la j^{e} classe (si H_0 est vraie). On a: $m_j = np_j$. De plus, on sait que la distribution binomiale tend vers la distribution normale. Finalement, le carré d'une distribution normale centrée réduite est une distribution du khi-deux à un degré de liberté (voir la page 226).

(iv) Comme on l'a mentionné au chapitre 5, les valeurs de $\chi^2_{\alpha,n}$ peuvent être obtenues en utilisant une calculatrice ou trouvées dans une table statistique (voir le tableau 5.3, page 240). On peut aussi se servir des formules d'approximation de Wilson-Hilferty et de Fisher (voir la page 240).

Exemple 6.2.1. On veut tester l'hypothèse que la durée de vie X (en mois) d'un composant électronique présente une distribution exponentielle. On a recueilli les données suivantes:

Intervalle j	$[0,10)$	$[10,20)$	$[20,30)$	$[30,40)$	$[40,\infty)$	Σ
Effectif n_j	35	20	18	8	19	100

De plus, la moyenne des données est de 25. On veut donc tester

$$H_0\text{: } X \sim \text{Exp}(\lambda)$$

contre

$$H_1\text{: } X \text{ ne présente } \textit{pas} \text{ une distribution exponentielle}$$

Supposons que $\alpha = 0{,}05$. On doit d'abord estimer le paramètre inconnu λ. On a déjà vu (voir la page 222) que l'estimateur à vraisemblance maximale de λ est donné par $1/\bar{X}$. Alors on pose (ce faisant, perdant un degré de liberté):

$$f_X(x) = \frac{1}{25} e^{-x/25} \quad \text{pour } x \geq 0$$

Maintenant, on trouve que

$$P[a \leq X < b] = e^{-a/25} - e^{-b/25} \quad \text{pour } 0 \leq a < b \leq \infty$$

d'où l'on déduit le tableau suivant:

j	$[0,10)$	$[10,20)$	$[20,30)$	$[30,40)$	$[40,\infty)$	Σ
n_j	35	20	18	8	19	$n = 100$
$(\simeq)\ p_j$	0,33	0,22	0,15	0,10	0,20	1
$(np_j =)\ m_j$	33	22	15	10	20	100

Ensuite, étant donné que $m_j \geq 5\ \forall j$, on calcule

$$d^2 = \frac{(35-33)^2}{33} + \ldots + \frac{(19-20)^2}{20} \simeq 1{,}35$$

Puisque $\chi^2_{0,05;5-1-1} \overset{\text{tab. 5.3}}{\simeq} 7{,}81$, on accepte le modèle exponentiel, avec $\lambda = 1/25$, au seuil $\alpha = 0{,}05$.

Remarques.

(i) Si l'on avait voulu tester, par exemple, l'hypothèse que la variable aléatoire X présente une distribution (approximativement) normale, alors il aurait fallu ajouter dans le tableau l'intervalle $(-\infty, 0)$ (dans lequel il n'y a naturellement aucune observation).

(ii) Comme nous l'avons mentionné auparavant, la valeur de β n'est pas unique. Dans l'exemple ci-dessus, si l'hypothèse H_1 est vraie, alors la variable aléatoire X peut présenter n'importe quelle distribution, sauf la distribution exponentielle. À chacune de ces distributions possibles, on peut associer une valeur de β.

(iii) Un test généralement plus puissant que celui de Pearson est le test de Kolmogorov-Smirnov. Pour tester la normalité, le test de Shapiro-Wilk (pour $n \leq 50$) est considéré par plusieurs comme le meilleur de tous. Nous étudierons les tests de Kolmogorov-Smirnov et de Shapiro-Wilk dans cette section. $\diamondsuit$

6.2.2 Test de Shapiro-Wilk

Soit $x_1, \ldots, x_n$ un échantillon aléatoire particulier d'une certaine population X. Pour faire le test de

$$H_0\text{: les données proviennent d'une population normale} \tag{6.10}$$

Shapiro et Wilk, dans un article publié en 1965 dans la revue scientifique *Biometrika* (voir la référence [27]), ont proposé une procédure qui est disponible dans de nombreux logiciels statistiques (dont le logiciel *SAS*, en particulier). Nous allons en donner ici les grandes lignes afin que les personnes utilisant ce test à l'aide d'un logiciel comprennent la procédure en question.
(i) On réarrange les données en ordre croissant; c'est-à-dire que l'on calcule les statistiques d'ordre $x_{(1)}, \ldots, x_{(n)}$, où $x_{(i)} \leq x_{(j)}$ si $i < j$.
(ii) On calcule

$$t^2 := \sum_{i=1}^{n} (x_{(i)} - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{6.11}$$

(iii) On calcule

$$b := \sum_{i=1}^{k} a_{n-i+1} (x_{(n-i+1)} - x_{(i)}) \tag{6.12}$$

où $k = n/2$ si n est pair et $k = (n-1)/2$ si n est impair. Les a_{n-i+1} sont des coefficients qui dépendent du vecteur des moyennes et de la *matrice de covariance* du vecteur aléatoire $\mathbf{Y} := (Y_1, \ldots, Y_n)$, où $Y_1, \ldots, Y_n$ est un échantillon aléatoire *ordonné* de taille n d'une population normale centrée réduite. La matrice de covariance du vecteur $\mathbf{Y}$ est la matrice $\mathbf{K}$ dont l'élément qui se trouve à l'intersection de la i^{e} ligne et de la j^{e} colonne est la covariance des variables aléatoires Y_i et Y_j. C'est-à-dire que

$$\mathbf{K} = (\sigma_{i,j}), \quad \text{où } \sigma_{i,j} := \text{COV}[Y_i, Y_j] \text{ pour } i, j = 1, \ldots, n \tag{6.13}$$

Shapiro et Wilk ont donné, dans leur article, une table des coefficients a_{n-i+1} pour $n = 2, \ldots, 50$. Si $n = 2k + 1$, c'est-à-dire si n est impair, alors on a: $a_{k+1} = 0$.

(iv) On calcule

$$W := \frac{b^2}{t^2} \tag{6.14}$$

(v) Il y a une table des *valeurs critiques* de W pour $n = 3, 4, \ldots, 50$ et plusieurs valeurs de α dans l'article de Shapiro et Wilk; c'est-à-dire les valeurs auxquelles la statistique W doit être comparée pour prendre une décision. Les valeurs de W sont situées entre (un peu moins de) 0,7 et 1. On rejette l'hypothèse nulle H_0 si la valeur de W est petite.

Exemple 6.2.2. Soit $x_1 = -2$, $x_2 = 3$, $x_3 = -3$, $x_4 = 5$, $x_5 = -1$, $x_6 = 0$ et $x_7 = -4$ un échantillon aléatoire particulier tiré d'une variable aléatoire X. On veut tester

$$H_0\text{: } X \sim \mathrm{N}(\mu, \sigma^2)$$

où μ et σ^2 sont des paramètres inconnus.

(i) On a: $x_{(1)} = -4$, $x_{(2)} = -3$, $x_{(3)} = -2$, $x_{(4)} = -1$, $x_{(5)} = 0$, $x_{(6)} = 3$ et $x_{(7)} = 5$.

(ii) On calcule

$$t^2 = \sum_{i=1}^{7} (x_{(i)} - \bar{x})^2 = \sum_{i=1}^{7} x_{(i)}^2 - 7\bar{x}^2 = 64 - \frac{4}{7} = \frac{444}{7}$$

On trouve dans la table préparée par Shapiro et Wilk que

$$a_{7-1+1} \simeq 0{,}6233, \quad a_{7-2+1} \simeq 0{,}3031, \quad a_{7-3+1} \simeq 0{,}1401, \quad a_{7-4+1} = 0$$

Il s'ensuit que

$$b \simeq (0{,}6233)(5 - (-4)) + (0{,}3031)(3 - (-3)) + (0{,}1401)(0 - (-2)) = 7{,}7085$$

(iv) On a:

$$W \simeq \frac{(7{,}7085)^2}{444/7} \simeq 0{,}937$$

(v) Puisque $0{,}937 > 0{,}928$, où (selon la table dans [27])

$$P[W \leq 0{,}928] \simeq 0{,}5$$

lorsque $n = 7$, on ne peut pas rejeter la normalité avec une valeur de α inférieure à 0,5. ♢

6.2.3 Test de Kolmogorov-Smirnov

Comme le test d'ajustement de Pearson, le test de Kolmogorov-Smirnov peut être utilisé pour n'importe quelle distribution. En pratique, il est surtout utilisé dans le cas où la variable aléatoire X est continue. On peut s'en servir, en particulier, pour tester la normalité (au lieu du test de Shapiro-Wilk) lorsque le nombre d'observations est supérieur à 50. Ce test est aussi disponible dans de nombreux logiciels statistiques.

On définit la **fonction de répartition de l'échantillon** ou **fonction de répartition empirique** comme suit:

$$F_n(x) = \begin{cases} 0 \text{ si } x < x_{(1)} \\ \frac{i}{n} \text{ si } x_{(i)} \leq x < x_{(i+1)}, \text{ pour } i = 1, \ldots, n-1 \\ 1 \text{ si } x \geq x_{(n)} \end{cases} \tag{6.15}$$

où $x_{(1)}, \ldots, x_{(n)}$ est un échantillon aléatoire particulier, rangé en ordre croissant, d'une variable aléatoire X. On pose:

$$D_n = \max_{x \in \mathbb{R}} |F_n(x) - F(x)| \tag{6.16}$$

où F est la fonction que l'on propose comme fonction de répartition théorique de X. Le test de

$$H_0\text{: la fonction de répartition de } X \text{ est } F \tag{6.17}$$

contre

$$H_1\text{: la fonction de répartition de } X \text{ n'est pas } F \tag{6.18}$$

proposé par Kolmogorov et Smirnov consiste à rejeter H_0 si D_n est plus grand qu'une certaine constante $D_{\alpha,n}$ que l'on peut trouver dans des tables statistiques (voir la référence [20] pour la table originale). De plus, lorsque H_0 est vraie, la distribution de D_n ne dépend pas de F et on peut donc utiliser la même table de valeurs critiques $D_{\alpha,n}$ pour toutes les fonctions F. Quelques valeurs de $D_{\alpha,n}$ sont présentées dans le tableau 6.1. Pour $n > 40$, les constantes $D_{\alpha,n}$ sont données approximativement par

$$\begin{cases} 1{,}63/\sqrt{n} \text{ si } \alpha = 0,01 \\ 1{,}36/\sqrt{n} \text{ si } \alpha = 0{,}05 \\ 1{,}22/\sqrt{n} \text{ si } \alpha = 0{,}10 \end{cases} \tag{6.19}$$

Tableau 6.1. Valeurs critiques de la statistique D_n

$n \setminus \alpha$	0,01	0,05	0,10
1	0,995	0,975	0,950
2	0,929	0,842	0,776
3	0,829	0,708	0,636
4	0,734	0,624	0,565
5	0,669	0,563	0,509
6	0,617	0,519	0,468
7	0,576	0,483	0,436
8	0,542	0,454	0,410
9	0,513	0,430	0,387
10	0,489	0,409	0,369
15	0,404	0,338	0,304
20	0,352	0,294	0,265
25	0,317	0,264	0,238
30	0,290	0,242	0,218
35	0,269	0,224	0,202
40	0,252	0,210	0,189

Enfin, on a:

$$P\left[D_n < \frac{d}{n}\right] \simeq 1 - 2e^{-2d^2} \quad \text{si } n \text{ est grand} \tag{6.20}$$

Remarque. Les valeurs de $D_{\alpha,n}$ que l'on trouve dans des tables statistiques ont été calculées pour le cas où la variable aléatoire X est continue. Si X est discrète, alors l'utilisation de la valeur critique $D_{\alpha,n}$ implique que le test correspondant possédera une erreur de première espèce qui est inférieure ou égale au α donné.

Cas particulier. Pour tester l'hypothèse nulle

$$H_0 \colon X \sim \mathrm{N}(\mu, \sigma^2) \tag{6.21}$$

on utilise la statistique

$$D_n = \max_{x \in \mathbb{R}} \left| F_n(x) - \Phi\left(\frac{x-\mu}{\sigma}\right) \right| \tag{6.22}$$

où Φ est la fonction de répartition d'une variable aléatoire $\mathrm{N}(0,1)$. On rejette H_0 si et seulement si la valeur de la statistique D_n est supérieure à $D_{\alpha,n}$.

Remarque. Dans le cas continu, on peut écrire que

$$D_n = \max_{1\le i\le n} \left\{ |F_n(x_i) - F(x_i)|, |F_n(x_i^-) - F(x_i)| \right\}$$

où $F_n(x_i^-) := \lim_{x\uparrow x_i} F_n(x)$ (la *limite à gauche* de $F_n(x)$ au point x_i). Puisque F_n est une fonction étagée, on a en fait simplement $F_n(x_i^-) = F_n(x_{i-1})$ si $i = 2, 3, \ldots, n$.

Exemple 6.2.3. (Voir la référence [18].) Soit

31,0 31,4 33,3 33,4 33,5 33,7 34,4 34,9 36,2 37,0

un échantillon aléatoire particulier de taille 10 d'une variable aléatoire X. On veut tester

$$H_0\text{: } X \text{ présente une distribution } \mathrm{N}(32; (1{,}8)^2)$$

contre

$$H_1\text{: } X \text{ ne présente pas une distribution } \mathrm{N}(32; (1{,}8)^2)$$

Remarque. Contrairement au test d'ajustement de Pearson, on ne peut pas tester simplement l'hypothèse que X présente une distribution normale; il faut préciser ses paramètres.

Considérons la statistique

$$\begin{aligned} d_{10} &= \max_{1\le i\le 10} \left\{ \left| F_{10}(x_i) - \Phi\left(\frac{x_i - 32}{1{,}8}\right)\right|, \left| F_{10}(x_i^-) - \Phi\left(\frac{x_i - 32}{1,8}\right)\right| \right\} \\ &:= \max_{1\le i\le 10} \{d_i, d_i^-\} \end{aligned}$$

On a:

x_i	$F_{10}(x_i)$	$\Phi((x_i - 32)/1{,}8)$	d_i	d_i^-
31,0	0,1	0,2895	0,189	0,289
31,4	0,2	0,3707	0,171	0,271
33,3	0,3	0,7642	0,464	0,564
33,4	0,4	0,7814	0,381	0,481
33,5	0,5	0,7967	0,297	0,397
33,7	0,6	0,8264	0,226	0,326
34,4	0,7	0,9082	0,208	0,308
34,9	0,8	0,9463	0,146	0,246
36,2	0,9	0,9901	0,090	0,190
37,0	1,0	0,9973	0,003	0,097

Donc, on peut écrire que $d_{10} \simeq 0{,}564$.

Maintenant, on trouve dans le tableau 6.1 des valeurs critiques de D_n que

$$P[D_{10} > 0{,}409] \simeq 0{,}05$$

Puisque $0{,}564 > 0{,}409$, on peut rejeter l'hypothèse nulle H_0 au seuil de signification $\alpha = 0{,}05$. ♢

Remarque. Lorsque n et α sont fixés, on cherche immédiatement la valeur critique de D_n dans le tableau. On peut s'arrêter et rejeter H_0 dès que l'on a obtenu une quantité $|F_n(x_i) - \Phi((x_i - \mu)/\sigma)|$ ou $|F_n(x_i^-) - \Phi((x_i - \mu)/\sigma)|$ supérieure à cette valeur critique.

6.3 Test d'indépendance

Supposons que les membres d'une certaine population peuvent être classés selon deux caractéristiques. Ces caractéristiques peuvent être qualitatives ou quantitatives; par exemple, le classement peut se faire selon le niveau d'études et le revenu annuel des gens. On désire tester l'hypothèse que les deux caractéristiques en question sont *indépendantes* pour les membres de cette population.

Soit X (respectivement Y) la valeur prise par la première (respectivement seconde) caractéristique. Supposons que X possède r classes (ou valeurs possibles) A_i et que Y possède s classes B_j. On veut donc tester:

$$P[X = A_i, Y = B_j] = P[X = A_i]P[Y = B_j] \quad \text{pour } i = 1, \ldots, r; j = 1, \ldots, s \tag{6.23}$$

On prélève un échantillon aléatoire de taille n de la population et on construit le tableau suivant (appelé **table de contingence**):

$X \setminus Y$	B_1	B_2	$\cdots$	B_s	
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	n_{1+}
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	n_{r+}
	n_{+1}	n_{+2}	$\cdots$	n_{+s}	n

où

$$\begin{aligned}
&n_{ij} && \text{est l'effectif de } (A_i, B_j) \text{ pour } i = 1, \ldots, r; j = 1, \ldots, s \\
&n_{i+} := \sum_{j=1}^{s} n_{ij} && \text{est l'effectif de } A_i \text{ pour } i = 1, \ldots, r \\
&n_{+j} := \sum_{i=1}^{r} n_{ij} && \text{est l'effectif de } B_j \text{ pour } j = 1, \ldots, s \\
&n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} && \text{est l'effectif total}
\end{aligned}$$

On calcule ensuite la statistique

$$D^2 := \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{6.24}$$

où

$$e_{ij} := \frac{(n_{i+})(n_{+j})}{n} \tag{6.25}$$

est la valeur que n_{ij} devrait prendre, en moyenne, si les variables X et Y sont indépendantes. On peut montrer que, si n est suffisamment grand et si l'hypothèse H_0 est vraie, alors D^2 présente approximativement une distribution du khi-deux à $(r-1)(s-1)$ degrés de liberté. Le test consiste à rejeter H_0 au seuil de signification α si et seulement si

$$D^2 > \chi^2_{\alpha,(r-1)(s-1)} \tag{6.26}$$

Remarques.
(i) Comme pour le cas du test d'ajustement de Pearson, il est préférable que e_{ij} soit supérieur ou égal à 5 pour tout couple (i, j).
(ii) Si $r = s = 2$, alors on devrait utiliser (la correction de continuité de Yates)

$$D^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|n_{ij} - e_{ij}| - \frac{1}{2})^2}{e_{ij}} \tag{6.27}$$

(iii) Ce test peut être employé pour vérifier si k proportions théoriques, $p_1, \ldots, p_k$, sont égales (voir la section 6.4, page 322).

Exemple 6.3.1. On s'intéresse à la durée de vie X (en milliers de kilomètres) des pneus de quatre marques différentes: A, B, C et D. Soit Y la marque des pneus. On a mesuré la durée de vie de 200 pneus de chaque marque et l'on a divisé cette durée de vie en trois classes: de 0 à 20 (milliers de kilomètres), de 20 à 30, et plus de 30. Les données sont les suivantes:

$X \setminus Y$	A	B	C	D	
$[0,20)$	26	23	15	32	96
$[20,30)$	118	93	116	121	448
$[30,\infty)$	56	84	69	47	256
	200	200	200	200	800

On veut tester l'indépendance des deux variables. On calcule

$$d^2 = \frac{(26-24)^2}{24} + \frac{(23-24)^2}{24} + \ldots + \frac{(47-64)^2}{64} \simeq 22{,}82$$

Puisque $D^2 \approx \chi^2_{(3-1)(4-1)} \equiv \chi^2_6$ (si H_0 est vraie) et $\chi^2_{0,05;6} \overset{\text{tab. 5.3}}{\simeq} 12{,}59$, on peut rejeter l'hypothèse d'indépendance au seuil de signification $\alpha = 0{,}05$.

Remarques.
(i) Ici, on a: $e_{ij} \geq 24 > 5\ \forall i, j$.
(ii) La variable *numérique* X devient une variable *qualitative* une fois que l'on a divisé l'ensemble de ses valeurs possibles en trois sous-ensembles comme ci-dessus. En effet, on pourrait dire que la durée de vie d'un pneu est classée selon l'une des trois catégories: faible, moyenne, élevée. ♢

6.4 Tests au sujet des paramètres

(I) Comparaisons entre un paramètre et une constante

6.4.1 Test d'une moyenne théorique μ; σ connu

Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n d'une variable aléatoire X de moyenne μ inconnue, mais variance σ^2 connue. On veut faire le test de

$$H_0\colon \mu = \mu_0 \quad \text{contre} \quad H_1\colon \mu \neq \mu_0 \tag{6.28}$$

Remarque. Le test ci-dessus est dit **bilatéral**. Un test **unilatéral à droite** (respectivement **à gauche**) est donné par

$$H_0\colon \mu = \mu_0 \quad \text{contre} \quad H_1\colon \mu > \mu_0 \quad (\text{respectivement } \mu < \mu_0) \tag{6.29}$$

Dans ce cas, on peut écrire l'hypothèse nulle H_0 comme suit:

$$H_0\colon \mu \leq \mu_0 \quad (\text{respectivement } \mu \geq \mu_0) \tag{6.30}$$

Cependant, dans ce manuel nous écrirons toujours l'hypothèse nulle sous la forme d'une hypothèse simple.

Pour effectuer le test en (6.28), on utilise la statistique

$$Z_0 := \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{6.31}$$

Lorsque H_0 est vraie, $\bar{X}$ présente une distribution normale (au moins approximativement, si n est suffisamment grand) de paramètres μ_0 et σ^2/n. Il s'ensuit que

$$Z_0 \overset{H_0}{\sim} \mathrm{N}(0,1) \tag{6.32}$$

Remarque. La notation $\overset{H_0}{\sim}$ signifie que Z_0 présente une distribution $\mathrm{N}(0,1)$ *si l'hypothèse H_0 est vraie.*

Le test consiste à rejeter H_0 au seuil de signification α si et seulement si

$$|Z_0| > z_{\alpha/2} \tag{6.33}$$

où $z_{\alpha/2}$ est tel que $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ (voir la section 5.4, page 233).

Remarques.
(i) On a bien:

$$P[\text{Rejeter } H_0 \mid H_0 \text{ est vraie}] = P[|Z_0| > z_{\alpha/2} \mid \mu = \mu_0] = 2(\alpha/2) = \alpha \tag{6.34}$$

C'est-à-dire que la **valeur critique** $z_{\alpha/2}$ est choisie de façon à obtenir un test de seuil α.
(ii) Dans le cas des tests unilatéraux, on rejette H_0 si et seulement si

$$\begin{cases} Z_0 > z_\alpha & \text{si } H_1\colon \mu > \mu_0 \\ Z_0 < -z_\alpha & \text{si } H_1\colon \mu < \mu_0 \end{cases} \tag{6.35}$$

(iii) On peut aussi écrire que l'on rejette l'hypothèse nulle au seuil α si et seulement si

$$\bar{X} < C_1 \quad \text{ou} \quad \bar{X} > C_2 \tag{6.36}$$

dans le cas du test bilatéral, où les **constantes de rejet** sont données par

$$C_1 = \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{et} \quad C_2 = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.37}$$

De même, on rejette H_0 si et seulement si

$$\begin{cases} \bar{X} > C := \mu_0 + z_\alpha \dfrac{\sigma}{\sqrt{n}} \text{ si } H_1\text{: } \mu > \mu_0 \\ \bar{X} < C := \mu_0 - z_\alpha \dfrac{\sigma}{\sqrt{n}} \text{ si } H_1\text{: } \mu < \mu_0 \end{cases} \tag{6.38}$$

(iv) On déduit de la remarque précédente qu'un intervalle de confiance bilatéral à $100(1-\alpha)$ % pour la moyenne μ de la variable aléatoire X qui présente une distribution $\mathrm{N}(\mu, \sigma^2)$ est en fait la région d'acceptation du test de H_0: $\mu = \mu_0$ contre H_1: $\mu \neq \mu_0$ au seuil de signification α. En effet, si l'intervalle de confiance ne contient pas le point μ_0, alors on peut rejeter H_0 au seuil α. De même, les intervalles de confiance unilatéraux sont les régions d'acceptation des tests d'hypothèses unilatéraux au sujet de μ.

Exemple 6.4.1. Soit $x_1, \ldots, x_{25}$ un échantillon aléatoire particulier d'une population normale de variance $\sigma^2 = 4$. Supposons que $\bar{x} = 2{,}8$. Pour tester

$$H_0\text{: } \mu = 2 \quad \text{contre} \quad H_1\text{: } \mu \neq 2$$

on calcule

$$z_0 = \frac{2{,}8 - 2}{2/\sqrt{25}} = 2$$

On choisit alors α, le seuil de signification du test. Soit $\alpha = 0{,}02$; alors $z_{\alpha/2} = z_{0{,}01} \overset{\text{tab. 5.1}}{\simeq} 2{,}326$. Puisque $|2| \leq 2{,}326$, on ne rejette pas l'hypothèse nulle. Par contre, si l'on choisit $\alpha = 0{,}05$, alors on obtient: $z_{\alpha/2} = z_{0{,}025} \simeq 1{,}960$. Étant donné que $|2| > 1,960$, on rejette l'hypothèse H_0: $\mu = 2$ au seuil $\alpha = 0{,}05$. $\diamondsuit$

Remarques.
(i) Les logiciels donnent habituellement la valeur de α qui correspond à la statistique observée z_0. Cette valeur est appelée **valeur** P. Dans le cas de l'exemple précédent, on trouve que $z_{\alpha/2} = 2$ si et seulement si $\alpha \simeq 0{,}0455$. Ainsi, si l'on choisit un α inférieur à 0,0455, alors on ne peut pas rejeter H_0.

En général, il est difficile de donner une formule exacte pour la valeur P; cependant, dans le cas du test d'une moyenne avec σ connu, on montre facilement que

$$P = \begin{cases} 2[1 - \Phi(|z_0|)] & \text{si } H_1\colon \mu \neq \mu_0 \\ 1 - \Phi(z_0) & \text{si } H_1\colon \mu > \mu_0 \\ \Phi(z_0) & \text{si } H_1\colon \mu < \mu_0 \end{cases} \tag{6.39}$$

(ii) Rappelons que, dans le cas des tests au sujet des paramètres d'une variable aléatoire, l'hypothèse nulle est habituellement celle que l'on croit fausse. Par conséquent, on essaie de la rejeter. En réalité, il s'agit d'un choix que nous avons fait pour la formulation des hypothèses. Ce ne sont pas tous les auteurs qui sont d'accord avec ce choix. Certains préfèrent poser que H_0 est l'hypothèse que l'on considère comme acceptable. De toute façon, on peut toujours intervertir le rôle des paramètres α et β si l'on change la formulation des hypothèses. Ainsi, tester

$$H_0\colon \mu = \mu_0 \quad \text{contre} \quad H_1\colon \mu > \mu_0$$

au seuil de signification $\alpha = 0{,}05$ et avec une valeur de β égale à 0,20 si $\mu = \mu_1 > \mu_0$ est équivalent à tester

$$H_0\colon \mu = \mu_1 \quad \text{contre} \quad H_1\colon \mu < \mu_1$$

au seuil $\alpha = 0{,}20$ et avec $\beta(\mu = \mu_0 < \mu_1) = 0{,}05$. De plus, parfois on ne peut pas choisir soi-même les hypothèses.

Erreur de deuxième espèce

Si l'hypothèse H_1 est vraie, c'est-à-dire si $\mu = \mu_1 = \mu_0 + \Delta$, où $\Delta \neq 0$, alors $\bar{X} \sim \mathrm{N}(\mu_1, \sigma^2/n)$ et il s'ensuit que

$$Z_0 \sim \mathrm{N}\left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}, 1\right) \equiv \mathrm{N}\left(\frac{\Delta\sqrt{n}}{\sigma}, 1\right) \tag{6.40}$$

Donc, si $H_1\colon \mu \neq \mu_0$, on a:

$$\begin{aligned} \beta\ (= \beta(\Delta)) &= P[\text{Accepter } H_0 \mid H_0 \text{ est fausse}] \\ &= P[-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2} \mid \mu = \mu_0 + \Delta] \\ &= P\left[-z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma} \leq \mathrm{N}(0,1) \leq z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma}\right] \\ &= \Phi\left(z_{\alpha/2} - \tfrac{\Delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \tfrac{\Delta\sqrt{n}}{\sigma}\right) \end{aligned} \tag{6.41}$$

Remarque. Dans le cas des tests unilatéraux, on trouve que (voir la figure 6.1)

$$\beta(\Delta) = \begin{cases} \Phi\left(z_\alpha - \frac{\Delta\sqrt{n}}{\sigma}\right) \text{ si } H_1\text{: } \mu > \mu_0 \\ \\ \Phi\left(z_\alpha + \frac{\Delta\sqrt{n}}{\sigma}\right) \text{ si } H_1\text{: } \mu < \mu_0 \end{cases} \tag{6.42}$$

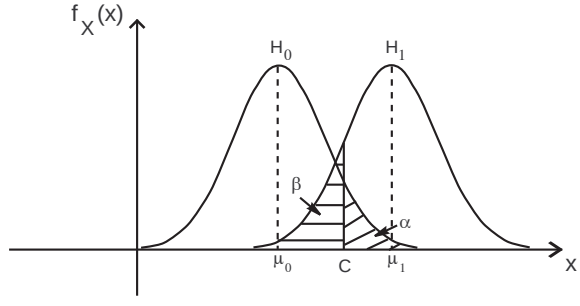

Fig. 6.1. Erreur de première et de deuxième espèce dans le cas du test unilatéral à droite

Exemple 6.4.1 (suite). Si la vraie valeur de la moyenne de la population est $\mu = \mu_1 = 2{,}6 = 2 + 0{,}6$, et si $\alpha = 0{,}05$, on calcule

$$\begin{aligned} \beta(0{,}6) &\simeq \Phi(1{,}960 - (0{,}6)(5/2)) - \Phi(-1{,}960 - (0{,}6)(5/2)) \\ &= \Phi(0{,}46) - \Phi(-3{,}46) \overset{\text{tab. A.3}}{\simeq} 0{,}6772 - (1 - 0{,}9997) = 0{,}6769 \end{aligned}$$

Donc la puissance du test, lorsque $\mu = 2{,}6$, est donnée par $1 - \beta(0{,}6) \simeq 0{,}3231$, ce qui est faible. Si l'on désire augmenter cette puissance, on peut recueillir plus d'observations. En effet, la valeur de β diminue lorsque n augmente (voir (6.41)). Par exemple, si l'on prend $n = 100$, alors on obtient que $\beta(0{,}6) \simeq 0{,}1492$, de sorte que $1 - \beta(0{,}6) \simeq 0{,}8508$. En pratique, on aimerait que $1 - \beta(\Delta)$ soit supérieur ou égal à 0,8 pour une différence Δ que l'on juge significative entre μ_0 et la vraie moyenne de la population. ♢

Taille de l'échantillon

On peut utiliser la formule (6.41) pour calculer la valeur de n requise pour obtenir un certain β, étant donné α et Δ. En effet, si $\Delta > 0$, alors on a:

$$\Phi\Big(-z_{\alpha/2} - \tfrac{\Delta\sqrt{n}}{\sigma}\Big) \simeq 0 \tag{6.43}$$

de sorte que

$$\beta \simeq \Phi\Big(z_{\alpha/2} - \tfrac{\Delta\sqrt{n}}{\sigma}\Big) \tag{6.44}$$

Cette dernière équation implique que

$$z_{1-\beta} \simeq z_{\alpha/2} - \frac{\Delta\sqrt{n}}{\sigma} \tag{6.45}$$

Puisque $z_{1-\beta} = -z_\beta$, on obtient la formule suivante:

$$n \simeq \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\Delta^2} \tag{6.46}$$

où $\Delta = \mu - \mu_0$.

Remarques.
(i) Si $\Delta < 0$, alors

$$\Phi\Big(z_{\alpha/2} - \tfrac{\Delta\sqrt{n}}{\sigma}\Big) \simeq 1 \tag{6.47}$$

et on obtient la même formule que ci-dessus pour n.
(ii) Comme la formule qui donne la valeur de β ne contient qu'un seul terme Φ si le test est unilatéral (voir (6.42)), on trouve le résultat *exact* suivant:

$$n = \frac{(z_\alpha + z_\beta)^2\sigma^2}{\Delta^2} \tag{6.48}$$

Exemple 6.4.2. Soit

$$H_0\colon \mu = 2 \quad \text{contre} \quad H_1\colon \mu \neq 2$$

Prenons $\alpha = 0{,}05$. Alors, si $\mu = 2{,}5$ et $\sigma^2 = 1$, pour que β soit inférieur ou égal à 0,10, il faut prendre

$$n \simeq \frac{(z_{0{,}025} + z_{0{,}1})^2 \cdot 1}{(2{,}5-2)^2} \overset{\text{tab. 5.1}}{\simeq} \frac{(1{,}960 + 1{,}282)^2}{0{,}25} \simeq 42$$

Remarque. Ici, on obtient:

$$\Phi\Big(-z_{\alpha/2} - \tfrac{\Delta\sqrt{n}}{\sigma}\Big) \simeq \Phi(-1{,}960 - (0{,}5)\sqrt{42}) \simeq \Phi(-5{,}20) \simeq 0$$

♢

6.4.2 Test d'une moyenne théorique μ; σ inconnu

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X qui présente une distribution normale de moyenne μ et variance σ^2 inconnues. Pour faire le test de

$$H_0\colon \mu = \mu_0 \tag{6.49}$$

on utilise la statistique

$$T_0 := \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \tag{6.50}$$

où S est l'écart-type de l'échantillon aléatoire. On peut écrire (voir la section 5.3, page 229) que

$$T_0 \overset{H_0}{\sim} t_{n-1} \tag{6.51}$$

Le test consiste à rejeter H_0 au seuil de signification α si et seulement si

$$\begin{cases} |T_0| > t_{\alpha/2,n-1} & \text{si } H_1\colon \mu \neq \mu_0 \\ T_0 > t_{\alpha,n-1} & \text{si } H_1\colon \mu > \mu_0 \\ T_0 < -t_{\alpha,n-1} & \text{si } H_1\colon \mu < \mu_0 \end{cases} \tag{6.52}$$

Exemple 6.4.3. Soit

$$5,\ 15,\ 14,\ 12,\ 4,\ 8,\ 9,\ 12,\ 10,\ 17,\ 12,\ 11$$

un échantillon aléatoire particulier d'une population normale de moyenne μ et variance σ^2 inconnues. On choisit $\alpha = 0{,}05$ et on veut tester

$$H_0\colon \mu = 15 \quad \text{contre} \quad H_1\colon \mu \neq 15$$

On trouve d'abord $\bar{x} = 10{,}75$ et $s \simeq 3{,}84$; ensuite, on calcule

$$t_0 \simeq \frac{10{,}75 - 15}{3{,}84/\sqrt{12}} \simeq -3{,}83$$

Finalement, puisque $t_{0{,}025;11} \overset{\text{tab. } 5.2}{\simeq} 2{,}201$, on a: $|t_0| > t_{0{,}025;11}$ et on peut rejeter l'hypothèse nulle H_0 au seuil $\alpha = 0{,}05$.

Remarque. On doit toujours poser les hypothèses H_0 et H_1 *avant* de prélever l'échantillon aléatoire. ♢

Erreur de deuxième espèce et taille de l'échantillon

Par définition, dans le cas bilatéral, on a:

$$\beta = P[\text{Accepter } H_0 \mid H_0 \text{ est fausse}] = P[|T_0| \leq t_{\alpha/2,n-1} \mid \mu = \mu_0 + \Delta] \quad (6.53)$$

où $\Delta \neq 0$. Maintenant, si $\mu = \mu_0 + \Delta$, alors on trouve que

$$T_0 := \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{Y}{\sqrt{W^2/(n-1)}} \quad (6.54)$$

où $Y \sim \mathrm{N}(\Delta\sqrt{n}/\sigma, 1)$ et $W^2 \sim \chi^2_{n-1}$ sont des variables aléatoires indépendantes. On dit que T_0 présente une distribution t *non centrale* à $n-1$ degrés de liberté et *paramètre de non-centralité* $\Delta\sqrt{n}/\sigma$. Par conséquent, il n'est pas facile de calculer la valeur *exacte* de β.

Il existe des formules d'approximation pour β. Par exemple, on peut montrer que

$$\beta(\mu) \simeq \Phi\left(\frac{t_{\alpha,n-1} - \delta\sqrt{n}}{\sqrt{1 + \frac{t^2_{\alpha,n-1}}{2n}}}\right) \quad \text{si } H_1\text{: } \mu > \mu_0 \quad (6.55)$$

où

$$\delta := \frac{\mu - \mu_0}{\sigma} \quad (6.56)$$

Puisque σ est inconnu, on doit le remplacer par son estimation ponctuelle s (si l'échantillon a déjà été prélevé) pour calculer δ; ou encore, on peut exprimer la différence $\mu - \mu_0$ en fonction de σ.

Il existe aussi des courbes ou abaques caractéristiques qui donnent la valeur de $\beta(\mu)$ en fonction d'un paramètre comme δ ci-dessus, pour un α fixé et pour plusieurs valeurs de n. On peut se servir de ces mêmes courbes pour calculer (approximativement) la taille n de l'échantillon à prélever pour obtenir une valeur donnée de β. Enfin, on trouve également des tableaux qui donnent la valeur de n correspondant à β, lorsque α et δ sont fixés.

Dans le cadre de ce manuel, pour le calcul de β ou de n, nous n'allons considérer généralement que le cas où la taille de l'échantillon (prélevé ou à prélever) est grande. On peut alors se ramener au cas où σ est connu et utiliser les formules déjà obtenues. Il faut cependant remplacer σ par s (si les données ont été recueillies).

6.4.3 Test d'une variance théorique σ^2

Supposons que $X_1, \ldots, X_n$ est un échantillon aléatoire d'une population normale X, de moyenne μ et variance σ^2 inconnues. Pour tester l'hypothèse nulle H_0: $\sigma^2 = \sigma_0^2$, on utilise la statistique

$$W_0^2 := \frac{(n-1)S^2}{\sigma_0^2} \tag{6.57}$$

Lorsque H_0 est vraie, W_0^2 présente une distribution du khi-deux à $n-1$ degrés de liberté. Le test consiste alors à rejeter H_0 au seuil de signification α si et seulement si

$$\begin{cases} W_0^2 > \chi^2_{\alpha/2,n-1} \quad \text{ou} \quad W_0^2 < \chi^2_{1-(\alpha/2),n-1} & \text{si } H_1\text{: } \sigma^2 \neq \sigma_0^2 \\ W_0^2 > \chi^2_{\alpha,n-1} & \text{si } H_1\text{: } \sigma^2 > \sigma_0^2 \\ W_0^2 < \chi^2_{1-\alpha,n-1} & \text{si } H_1\text{: } \sigma^2 < \sigma_0^2 \end{cases} \tag{6.58}$$

Remarques.
(i) On peut formuler le critère de rejet en fonction de S^2 comme suit: on rejette H_0 au seuil α si et seulement si

$$S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{\alpha,n-1} \quad \text{si } H_1\text{: } \sigma^2 > \sigma_0^2 \tag{6.59}$$

Il en est de même dans le cas des deux autres tests.

(ii) On a déjà mentionné, au chapitre précédent (voir la page 241), que si n est suffisamment grand, alors on peut écrire que

$$S \approx \mathrm{N}\Big(\sigma, \tfrac{\sigma^2}{2n}\Big) \tag{6.60}$$

Il s'ensuit que

$$Z_0 := \frac{S - \sigma_0}{\sigma_0/\sqrt{2n}} \overset{H_0}{\approx} \mathrm{N}(0,1) \tag{6.61}$$

Ainsi, lorsque n est grand, on peut aussi rejeter H_0 si et seulement si

$$\begin{cases} |Z_0| > z_{\alpha/2} & \text{si } H_1\text{: } \sigma^2 \neq \sigma_0^2 \\ Z_0 > z_\alpha & \text{si } H_1\text{: } \sigma^2 > \sigma_0^2 \\ Z_0 < -z_\alpha & \text{si } H_1\text{: } \sigma^2 < \sigma_0^2 \end{cases} \tag{6.62}$$

(iii) Si l'on suppose que la moyenne μ de la population X est connue, alors on devrait remplacer la statistique W_0^2 par

$$W_{00}^2 := \frac{n\hat{\sigma}^2}{\sigma_0^2} \tag{6.63}$$

où $\hat{\sigma}^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$, et comparer W_{00}^2 aux quantiles d'une distribution du khi-deux à n degrés de liberté, puisque $W_{00}^2 \overset{H_0}{\sim} \chi_n^2$.

Exemple 6.4.4. (Voir la référence [4].) Soit

$$10{,}2 \quad 9{,}9 \quad 9{,}8 \quad 10{,}1 \quad 10{,}1 \quad 9{,}9 \quad 10{,}0 \quad 9{,}7 \quad 10{,}1 \quad 10{,}3$$

un échantillon aléatoire particulier d'une variable aléatoire X qui désigne le poids de boîtes de carton contenant du savon. On suppose que X présente approximativement une distribution $\mathrm{N}(\mu, \sigma^2)$ et on veut tester

$$H_0\colon \sigma^2 = 0{,}04 \quad \text{contre} \quad H_1\colon \sigma^2 > 0{,}04$$

au seuil $\alpha = 0{,}05$. On trouve que $\bar{x} = 10{,}01$ et $\sum_{i=1}^{10} x_i^2 = 1002{,}31$. Il s'ensuit que

$$s^2 = \sum_{i=1}^{10} \frac{(x_i - \bar{x})^2}{10-1} = \frac{1}{9}\left[\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right] \simeq \frac{0{,}31}{9}$$

de sorte que

$$w_0^2 \simeq \frac{0{,}31}{0{,}04} = 7{,}75$$

Finalement, étant donné que $\chi^2_{0{,}05;9} \overset{\text{tab. } 5.3}{\simeq} 16{,}92$, on ne rejette pas l'hypothèse nulle H_0 au seuil $\alpha = 0{,}05$.

Remarque. Ici, puisque $s^2 \simeq 0{,}034$ est inférieur à 0,04, on ne pouvait certainement pas rejeter H_0 avec un α petit. ♢

Erreur de deuxième espèce et taille de l'échantillon

Supposons que la contre-hypothèse est $H_1\colon \sigma^2 < \sigma_0^2$ et que la vraie valeur de σ^2 est $\sigma_1^2 < \sigma_0^2$. On a:

$$\begin{aligned}\beta &= P[W_0^2 \geq \chi^2_{1-\alpha,n-1} \mid \sigma^2 = \sigma_1^2] \\ &= P\left[\frac{\sigma_1^2}{\sigma_0^2}\frac{(n-1)S^2}{\sigma_1^2} \geq \chi^2_{1-\alpha,n-1} \mid \sigma^2 = \sigma_1^2\right] \\ &= P[\chi^2_{n-1} \geq \lambda\chi^2_{1-\alpha,n-1}] \end{aligned} \tag{6.64}$$

où

$$\lambda := \frac{\sigma_0^2}{\sigma_1^2} \tag{6.65}$$

De même, on trouve que

$$\beta(\lambda) = \begin{cases} P[\chi^2_{n-1} \leq \lambda\chi^2_{\alpha,n-1}] & \text{si } H_1\text{: } \sigma^2 > \sigma_0^2 \\ P[\lambda\chi^2_{1-(\alpha/2),n-1} \leq \chi^2_{n-1} \leq \lambda\chi^2_{\alpha/2,n-1}] & \text{si } H_1\text{: } \sigma^2 \neq \sigma_0^2 \end{cases} \tag{6.66}$$

Maintenant, si n est grand, alors on peut écrire (voir la page 226) que

$$\chi^2_n \approx \mathrm{N}(n, 2n) \tag{6.67}$$

On peut donc calculer approximativement $\beta(\lambda)$ à l'aide des formules précédentes.

Pour déterminer la taille de l'échantillon qui permet d'obtenir un β inférieur ou égal à une valeur donnée, on peut utiliser les formules suivantes:

$$n \simeq \begin{cases} \dfrac{3}{2} + \dfrac{1}{2}\left(\dfrac{\sigma_0 z_\alpha + \sigma_1 z_\beta}{\sigma_1 - \sigma_0}\right)^2 & \text{(cas unilatéraux)} \\ \dfrac{3}{2} + \dfrac{1}{2}\left(\dfrac{\sigma_0 z_{\alpha/2} + \sigma_1 z_\beta}{\sigma_1 - \sigma_0}\right)^2 & \text{(cas bilatéral)} \end{cases} \tag{6.68}$$

où σ_1^2 est la vraie variance de la population. On pourrait aussi se servir de ces formules pour trouver la valeur de β, étant donné n, α et σ_1.

Exemple 6.4.4 (suite). Supposons que l'on a recueilli 51 observations plutôt que 10 dans l'exemple précédent. Si $\sigma_1^2 = 0{,}06$, alors $\lambda = 2/3$ et on calcule

$$\begin{aligned}\beta(\lambda = 2/3) &= P[\chi^2_{50} \leq (2/3)\chi^2_{0{,}05;50}] \\ &\overset{\text{tab. 5.3}}{\simeq} P[\mathrm{N}(50, 100) \leq 45] = \Phi(-0{,}5) \overset{\text{tab. A.3}}{\simeq} 0{,}31\end{aligned}$$

Si l'on se sert de la première formule dans (6.68), on trouve que

$$51 = \frac{3}{2} + \frac{1}{2}\left(\frac{(0{,}2)(1{,}645) + \sqrt{0{,}06}z_\beta}{\sqrt{0{,}06} - 0{,}2}\right)^2$$

de sorte que

$$(\sqrt{0{,}06}z_\beta + 0{,}33)^2 \simeq 0{,}2 \implies z_\beta \simeq 0{,}48 \implies \beta \simeq 0{,}32$$

car $\Phi(0,48) \overset{\text{tab. A.3}}{\simeq} 0{,}68$.

Finalement, si l'on veut être capable de détecter avec une probabilité d'au moins 90 % que σ^2 est supérieur à 0,04, lorsque la variance de la population est en fait de 0,06, alors il faut prendre

$$n \simeq \frac{3}{2} + \frac{1}{2}\left(\frac{0{,}2z_{0,05} + \sqrt{0{,}06}z_{0,10}}{\sqrt{0{,}06} - 0{,}2}\right)^2 \overset{\text{tab. 5.1}}{\simeq} 103{,}83$$

Donc, il faut recueillir au moins 104 observations. ♢

6.4.4 Test d'une proportion théorique *p*

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une certaine population, et soit p la proportion des individus dans la *population* qui présentent une caractéristique donnée. On veut tester

$$H_0\colon p = p_0 \tag{6.69}$$

Soit X le nombre d'individus dans l'*échantillon aléatoire* de taille n qui possèdent la caractéristique considérée. On sait que $X \sim \mathrm{B}(n,p)$. Alors, par le théorème central limite, on peut écrire que $X \approx \mathrm{N}(np, np(1-p))$. Donc, si l'hypothèse nulle H_0 est vraie, alors $X \approx \mathrm{N}(np_0, np_0(1-p_0))$, de sorte que

$$Z_0 := \frac{X - np_0}{\sqrt{np_0(1-p_0)}} \overset{H_0}{\approx} \mathrm{N}(0,1) \tag{6.70}$$

si n est suffisamment grand. Le test consiste à rejeter H_0 au seuil de signification α si et seulement si

$$\begin{cases} |Z_0| > z_{\alpha/2} & \text{si } H_1\colon p \neq p_0 \\ Z_0 > z_\alpha & \text{si } H_1\colon p > p_0 \\ Z_0 < -z_\alpha & \text{si } H_1\colon p < p_0 \end{cases} \tag{6.71}$$

Remarques.
(i) On pourrait tenir compte de la correction de continuité (voir la page 90) lorsqu'on approche la distribution binomiale par la distribution normale.

(ii) Il existe une procédure, basée sur la distribution binomiale, pour tester H_0 lorsque la taille de l'échantillon est petite.

(iii) On a: $\hat{p} = X/n$. Alors le test peut aussi s'énoncer comme suit: on rejette H_0: $p = p_0$ au seuil de signification α si et seulement si

$$\begin{cases} \hat{p} \begin{cases} < C_1 := p_0 - z_{\alpha/2}\sqrt{\dfrac{p_0(1-p_0)}{n}} \\ \text{ou} \\ > C_2 := p_0 + z_{\alpha/2}\sqrt{\dfrac{p_0(1-p_0)}{n}} \end{cases} & \text{si } H_1\text{: } p \neq p_0 \\ \hat{p} > C := p_0 + z_\alpha\sqrt{\dfrac{p_0(1-p_0)}{n}} & \text{si } H_1\text{: } p > p_0 \\ \hat{p} < C := p_0 - z_\alpha\sqrt{\dfrac{p_0(1-p_0)}{n}} & \text{si } H_1\text{: } p < p_0 \end{cases} \tag{6.72}$$

Exemple 6.4.5. Soit p la proportion des électeurs favorables à un certain candidat dans une élection. On veut faire le test de

$$H_0\text{: } p = 0{,}6 \quad \text{contre} \quad H_1\text{: } p < 0{,}6$$

C'est-à-dire que l'on croit que la popularité du candidat est en fait inférieure à 60 % des électeurs. On interroge 100 personnes (qui ont le droit de vote) prises au hasard (et sans remise); 55 appuient le candidat en question. Alors on calcule

$$z_0 = \frac{55 - (100)(0{,}6)}{\sqrt{100(0{,}6)(1-0{,}6)}} \simeq -1{,}02$$

Puisque $-z_{0{,}05} \overset{\text{tab. } 5.1}{\simeq} -1{,}645$, on ne peut pas rejeter l'hypothèse nulle H_0 au seuil de signification $\alpha = 0{,}05$. Notons qu'ici le nombre X de personnes favorables au candidat dans l'échantillon doit être inférieur ou égal à 51 pour pouvoir rejeter H_0 au seuil $\alpha = 0{,}05$ (avec $X = 52$, on trouve que $z_0 \simeq -1{,}63$). ♢

Erreur de deuxième espèce et taille de l'échantillon

Il n'est pas difficile de développer des formules pour la valeur de β et pour la taille de l'échantillon requise pour obtenir un β donné. On trouve que si la contre-hypothèse est H_1: $p \neq p_0$, alors $\beta(p) \simeq$

$$\Phi\left(\frac{\sqrt{n}(p_0 - p) + z_{\alpha/2}\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{\sqrt{n}(p_0 - p) - z_{\alpha/2}\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}}\right) \tag{6.73}$$

Dans les cas unilatéraux, on a:

$$\beta(p) \simeq \begin{cases} \Phi\left(\dfrac{\sqrt{n}(p_0 - p) + z_\alpha\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}}\right) \text{ si } H_1\text{: } p > p_0 \\ \Phi\left(\dfrac{\sqrt{n}(p - p_0) + z_\alpha\sqrt{p_0(1-p_0)}}{\sqrt{p(1-p)}}\right) \text{ si } H_1\text{: } p < p_0 \end{cases} \tag{6.74}$$

Finalement, la formule pour la taille de l'échantillon est

$$n \simeq \frac{1}{(p-p_0)^2}\left(z_{\alpha/2}\sqrt{p_0(1-p_0)} + z_\beta\sqrt{p_1(1-p_1)}\right)^2 \tag{6.75}$$

dans le cas bilatéral; dans les deux cas unilatéraux, il suffit de remplacer $\alpha/2$ par α dans la formule.

(II) Comparaisons entre deux paramètres

6.4.5 Test de l'égalité de deux moyennes; variances connues

Soit $X_1, \ldots, X_m$ un échantillon aléatoire d'une variable aléatoire X dont la moyenne μ_X est inconnue et la variance σ_X^2 est connue, et $Y_1, \ldots, Y_n$ un échantillon aléatoire d'une variable aléatoire Y de moyenne μ_Y inconnue et variance σ_Y^2 connue. On suppose que les X_i et les Y_j sont des variables aléatoires indépendantes $\forall i, j$. Pour faire le test de

$$H_0\text{: } \mu_X - \mu_Y = \Delta \tag{6.76}$$

on utilise la statistique

$$Z_0 := \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \tag{6.77}$$

Puisque $\bar{X}$ et $\bar{Y}$ sont indépendantes, on a:

$$\bar{X} - \bar{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right) \tag{6.78}$$

(au moins approximativement si m et n sont suffisamment grands). Il s'ensuit que

$$Z_0 \overset{H_0}{\sim} \mathrm{N}(0,1) \tag{6.79}$$

On rejette alors H_0 au seuil de signification α si et seulement si

$$\begin{cases} |Z_0| > z_{\alpha/2} \text{ si } H_1\text{: } \mu_X - \mu_Y \neq \Delta \\ Z_0 > z_\alpha \quad \text{ si } H_1\text{: } \mu_X - \mu_Y > \Delta \end{cases} \tag{6.80}$$

Remarques.
(i) Dans le cas des tests où l'on compare deux paramètres, on peut se limiter aux deux contre-hypothèses ci-dessus puisque l'on veut déterminer si les deux paramètres diffèrent d'une constante Δ (H_0) ou non (H_1), ou bien si la différence entre *un* paramètre et *l'autre* est inférieure ou égale à Δ (H_0) ou supérieure à Δ (H_1).
(ii) Comme nous l'avons mentionné dans le cas du test d'une moyenne théorique μ (avec σ connu), à la page 301, un intervalle de confiance peut être considéré comme la région d'acceptation d'un test d'hypothèses. Par exemple, si l'on construit un intervalle de confiance bilatéral à 95 % pour la différence $\mu_X - \mu_Y$ et si cet intervalle ne contient pas le point 0, alors on peut rejeter l'hypothèse H_0: $\mu_X = \mu_Y$ et accepter H_1: $\mu_X \neq \mu_Y$ au seuil de signification $\alpha =$ 0,05. De même, si l'intervalle de confiance est unilatéral, avec borne inférieure, et si cette borne inférieure est positive, alors on peut conclure que l'hypothèse H_1: $\mu_X > \mu_Y$ est vraie.

Exemple 6.4.6. Soit $x_1, \ldots, x_{10}$ un échantillon aléatoire particulier de $X \sim \mathrm{N}(\mu_X, 2)$ et $y_1, \ldots, y_{15}$ un échantillon aléatoire particulier de $Y \sim \mathrm{N}(\mu_Y, 3)$. On suppose que les échantillons sont indépendants et on désire tester

$$H_0\text{: } \mu_X = \mu_Y \quad \text{contre} \quad H_1\text{: } \mu_Y > \mu_X$$

au seuil $\alpha = 0{,}05$. Il suffit de prendre $\Delta = 0$ et d'inverser le rôle de X et de Y dans les formules ci-dessus.

Supposons que $\bar{x} = 4$ et $\bar{y} = 5$; alors on calcule

$$z_0 = \frac{5-4}{\sqrt{\frac{3}{15} + \frac{2}{10}}} \simeq 1{,}58$$

Puisque $z_{0,05} \overset{\text{tab. 5.1}}{\simeq} 1{,}645$, on ne peut pas rejeter l'hypothèse nulle H_0: $\mu_X = \mu_Y$ au seuil $\alpha = 0{,}05$.

Remarque. Si l'on avait calculé $z_0 = 1{,}68$, par exemple, alors on pourrait rejeter H_0 et accepter H_1: $\mu_X < \mu_Y$. Cependant, on ne pourrait pas rejeter H_0 si la contre-hypothèse était plutôt H_1: $\mu_X \neq \mu_Y$, car $|1{,}68| = 1{,}68$ n'est pas plus grand que $z_{0,025} \overset{\text{tab. 5.1}}{\simeq} 1{,}960$. Dans un tel cas, cela signifie que l'on devrait recueillir d'autres observations afin d'arriver à une conclusion plus forte. ♢

Erreur de deuxième espèce et taille des échantillons

En procédant comme dans le cas du test d'une moyenne théorique μ avec σ connu, on trouve que

$$\beta(\delta) = \begin{cases} \Phi(z_{\alpha/2} - \delta) - \Phi(-z_{\alpha/2} - \delta) & \text{si } H_1\text{: } \mu_X - \mu_Y \neq \Delta \\ \Phi(z_\alpha - \delta) & \text{si } H_1\text{: } \mu_X - \mu_Y > \Delta \end{cases} \tag{6.81}$$

où

$$\delta := \frac{\mu_X - \mu_Y - \Delta}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \tag{6.82}$$

De même, on trouve que, pour obtenir un β donné, il faut prendre des échantillons de taille

$$m = n \simeq \frac{(z_{\alpha/2} + z_\beta)^2(\sigma_X^2 + \sigma_Y^2)}{(\mu_X - \mu_Y - \Delta)^2} \tag{6.83}$$

lorsque la contre-hypothèse est H_1: $\mu_X - \mu_Y \neq \Delta$. Dans le cas du test unilatéral, on remplace $\alpha/2$ par α dans la formule précédente, et le résultat n'est plus seulement approximatif mais exact.

Remarque. En théorie, il n'est pas nécessaire de prendre $m = n$ $(= c)$. On peut choisir m, par exemple, et alors la valeur de n qui correspond à un β donné est $2c - m$. On peut procéder ainsi, en particulier, lorsqu'il est plus coûteux de prélever un échantillon d'une variable aléatoire que de l'autre.

6.4.6 Test de l'égalité de deux moyennes; variances inconnues

Soit $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$ des échantillons aléatoires d'une variable aléatoire X et d'une variable aléatoire Y, respectivement. On suppose que tous les paramètres sont inconnus et que X et Y présentent des distributions normales indépendantes. On veut tester

$$H_0\colon \mu_X - \mu_Y = \Delta \tag{6.84}$$

1^er^ cas: $\sigma_X^2 = \sigma_Y^2$

On suppose, d'abord, que les variances sont inconnues mais égales. En pratique, on effectue un test de l'égalité des variances (que nous verrons à la page 319) pour s'assurer que cette hypothèse est plausible. On définit

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2} \tag{6.85}$$

où S_X^2 et S_Y^2 sont les variances des échantillons aléatoires, et on considère la statistique

$$T_0 := \frac{\bar{X} - \bar{Y} - \Delta}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \tag{6.86}$$

On peut écrire (voir la page 238) que

$$T_0 \overset{H_0}{\sim} t_{m+n-2} \tag{6.87}$$

Il s'ensuit que l'on rejette H_0 au seuil de signification α si et seulement si

$$\begin{cases} |T_0| > t_{\alpha/2,m+n-2} & \text{si } H_1\colon \mu_X - \mu_Y \neq \Delta \\ T_0 > t_{\alpha,m+n-2} & \text{si } H_1\colon \mu_X - \mu_Y > \Delta \end{cases} \tag{6.88}$$

Exemple 6.4.7. Supposons que l'on désire comparer le temps d'utilisation (en minutes) de deux marques, A et B, de piles pour ordinateurs portables. On a les informations suivantes obtenues à partir de deux échantillons aléatoires particuliers de tailles égales à 25:

$$\bar{x}_A \simeq 181{,}08, \quad s_A \simeq 2{,}465, \quad \bar{x}_B \simeq 182{,}76, \quad s_B \simeq 1{,}877$$

On pourrait vérifier, à l'aide d'un test, l'hypothèse que les variances des populations dont proviennent les données peuvent être supposées égales. Soit X_A (respectivement X_B) le temps d'utilisation des piles de marque A (respectivement B). On suppose que $X_A \approx \mathrm{N}(\mu_A, \sigma^2)$ et $X_B \approx \mathrm{N}(\mu_B, \sigma^2)$, où tous les paramètres sont inconnus, et que X_A et X_B sont des variables aléatoires indépendantes. On veut tester

$$H_0\colon \mu_A = \mu_B \quad \text{contre} \quad H_0\colon \mu_A \neq \mu_B$$

au seuil de signification $\alpha = 0{,}05$. On calcule d'abord

$$s_p^2 \simeq \frac{24 \cdot (2{,}465)^2 + 24 \cdot (1{,}877)^2}{25 + 25 - 2} \simeq 4{,}80$$

Ensuite, on a:

$$t_0 \simeq \frac{181{,}08 - 182{,}76}{2{,}19\sqrt{\frac{1}{25} + \frac{1}{25}}} \simeq -2{,}71$$

Finalement, étant donné que $t_{0{,}025;48} < t_{0{,}025;40} \simeq 2{,}021$ et $|-2{,}71| > 2{,}021$, on peut rejeter l'égalité des moyennes au seuil $\alpha = 0{,}05$.

Remarquons que la différence entre les moyennes des échantillons est relativement petite. Cependant, les écarts-types des échantillons aussi sont petits, ce qui permet de conclure qu'il y a en fait une différence significative entre les moyennes des temps d'utilisation des piles de marque A et celles de marque B. On pourrait également conclure (avec $H_1\colon \mu_A < \mu_B$) que les piles de marque B durent plus longtemps, en moyenne, que celles de marque A avant de devoir être rechargées, et ce, avec un seuil α inférieur à 0,005. ♢

2e cas: $\sigma_X^2 \neq \sigma_Y^2$

Si l'on ne peut pas supposer que $\sigma_X^2 = \sigma_Y^2$, alors on doit utiliser la statistique

$$T_0^* := \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \tag{6.89}$$

qui, lorsque H_0 est vraie, présente *approximativement* une distribution t de Student à ν degrés de liberté, où

$$\nu := \frac{(a+b)^2}{\frac{a^2}{m-1} + \frac{b^2}{n-1}} \tag{6.90}$$

avec

$$a := \frac{S_X^2}{m} \quad \text{et} \quad b := \frac{S_Y^2}{n} \tag{6.91}$$

On remplace alors T_0 par T_0^* et $m+n-2$ par ν dans le cas précédent.

Remarques.
(i) On trouve aussi dans plusieurs livres la formule

$$\nu := \frac{(a+b)^2}{\frac{a^2}{m+1} + \frac{b^2}{n+1}} - 2 \tag{6.92}$$

(ii) On peut montrer que $\min\{m-1, n-1\} \leq \nu \leq m+n-2$. Cette formule est intéressante, car elle nous évite souvent d'avoir à calculer ν explicitement. En effet, si la conclusion du test est la même pour les deux valeurs extrêmes que ν peut prendre, alors il est inutile de connaître sa valeur exacte. De toute façon, la formule (6.90) ne donnera généralement pas un entier, de sorte qu'il faudra arrondir la valeur numérique obtenue.

(iii) Quand les tailles des échantillons sont assez grandes, on peut utiliser la distribution normale centrée réduite comme distribution approximative (sous H_0) de T_0 ou T_0^*.

(iv) Des formules pour la valeur de β et la taille des échantillons peuvent être obtenues, mais on se sert plutôt d'abaques caractéristiques en pratique.

Exemple 6.4.8. Soit $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ deux variables aléatoires indépendantes. On suppose que tous les paramètres sont inconnus. On prélève un échantillon aléatoire particulier de taille $n_1 = 4$ (respectivement $n_2 = 3$) de X_1 (respectivement X_2). On obtient:

$$\bar{x}_1 \simeq 291{,}25, \quad s_1^2 \simeq 190{,}25, \quad \bar{x}_2 \simeq 316{,}00, \quad s_2^2 \simeq 12{,}00$$

On veut tester l'égalité des moyennes au seuil de signification $\alpha = 0{,}10$. On voit que les variances des échantillons sont très différentes. Alors on effectue le test avec $\sigma_1^2 \neq \sigma_2^2$. On calcule

$$t_0^* \simeq \frac{291{,}25 - 316{,}00}{\sqrt{\frac{190{,}25}{4} + \frac{12{,}00}{3}}} \simeq -3{,}45$$

Le nombre ν de degrés de liberté de la statistique T_0^* sous H_0 est donné par

$$\nu \simeq \frac{(47{,}56+4)^2}{\frac{(47{,}56)^2}{3}+\frac{4^2}{2}} \simeq 3{,}49$$

Puisque $|-3{,}45| > t_{0{,}05;3} \simeq 2{,}353 > t_{0{,}05;4} \simeq 2{,}132$, on peut rejeter H_0 au seuil $\alpha = 0{,}10$.

Remarque. On a: $2 \leq \nu \leq 5$ et $t_{0{,}05;5} < t_{0{,}05;2} \simeq 2{,}920$. Étant donné que $|-3{,}45| > t_{0{,}05;2}$, il n'était pas nécessaire de calculer explicitement ν pour pouvoir conclure. ♢

6.4.7 Test de deux moyennes avec observations appariées

Supposons que l'on recueille des couples $(X_1, Y_1), \ldots, (X_n, Y_n)$ d'observations de variables aléatoires $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ et que les conditions sous lesquelles les observations sont recueillies varient (ou peuvent varier) d'un couple à l'autre. Dans ce cas, les observations X_k et Y_k ne sont *pas* indépendantes. Pour tester H_0: $\mu_X - \mu_Y = \Delta$, on définit la variable $D = X - Y$ et on calcule les différences $D_k := X_k - Y_k$ pour $k = 1, \ldots, n$. Le problème de tester H_0 est alors ramené à celui de faire le test d'une moyenne théorique μ_D avec σ_D inconnu, et on peut utiliser les formules du test d'une moyenne théorique μ avec σ inconnu (voir la page 305).

Exemple 6.4.9. Dans un test de dureté de matériaux, une boule d'acier est pressée contre chaque matériau et le diamètre de l'enfoncement est mesuré. Ce diamètre est relié à la dureté. On dispose de 2 types de boules d'acier, que l'on utilise avec 10 matériaux différents. On désire déterminer si les deux types d'acier donnent des résultats équivalents. On a recueilli les données suivantes:

Matériau	1	2	3	4	5	6	7	8	9	10
Acier 1	75	46	57	43	58	32	61	56	34	65
Acier 2	52	41	43	47	32	49	52	44	57	60

Soit X le diamètre de l'enfoncement obtenu avec l'acier 1 et Y le diamètre avec l'acier 2. On définit $D = X - Y$ et on veut tester

$$H_0\colon \mu_D = 0 \quad \text{contre} \quad H_1\colon \mu_D \neq 0$$

Soit $\alpha = 0{,}05$. On calcule les différences d_k pour $k = 1, \ldots, 10$:

Matériau	1	2	3	4	5	6	7	8	9	10
d_k	23	5	14	−4	26	−17	9	12	−23	5

On trouve que $\bar{d} = 5$ et que $s_D \simeq 15{,}85$. On calcule ensuite

$$t_0 := \frac{\bar{d}}{s_D/\sqrt{n}} \simeq \frac{5 \cdot \sqrt{10}}{15{,}85} \simeq 1$$

Puisque $t_{0{,}025;10-1} \overset{\text{tab. 5.2}}{\simeq} 2{,}262$, on ne rejette pas H_0 au seuil $\alpha = 0{,}05$. ♢

6.4.8 Test de l'égalité de deux variances

Soit $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ et $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ deux variables aléatoires indépendantes, et soit S_X^2 et S_Y^2 les variances d'un échantillon aléatoire de taille m de X et d'un échantillon aléatoire de taille n de Y, respectivement. On suppose que tous les paramètres sont inconnus. Pour tester l'hypothèse nulle

$$H_0\colon \sigma_X^2 = \sigma_Y^2 \tag{6.93}$$

on utilise la statistique

$$F_0 := \frac{S_X^2}{S_Y^2} \tag{6.94}$$

On peut montrer (voir la page 229) que si H_0 est vraie, alors F_0 présente une distribution F (de Fisher) à $m-1$ et $n-1$ degrés de liberté. C'est-à-dire que $F_0 \overset{H_0}{\sim} F_{m-1,n-1}$. On rejette H_0 au seuil de signification α si et seulement si

$$\begin{cases} F_0 > F_{\alpha/2,m-1,n-1} \quad \text{ou} \quad F_0 < F_{1-(\alpha/2),m-1,n-1} & \text{si } H_1\colon \sigma_X^2 \neq \sigma_Y^2 \\ F_0 > F_{\alpha,m-1,n-1} & \text{si } H_1\colon \sigma_X^2 > \sigma_Y^2 \end{cases} \tag{6.95}$$

Remarques.
(i) Rappelons que les valeurs critiques F_{α,n_1,n_2} sont définies de la même manière que les autres valeurs critiques ($z_\alpha, t_{\alpha,n}$ et $\chi^2_{\alpha,n}$):

$$P[F \leq F_{\alpha,n_1,n_2}] = 1 - \alpha \quad \text{si } F \sim F_{n_1,n_2} \tag{6.96}$$

(voir la section 5.4, page 242).

(ii) Comme dans le cas du test d'une variance théorique σ^2, un test approximatif, valide lorsque les tailles des échantillons sont grandes, peut être construit en se basant sur la distribution normale. En effet, si m et n sont assez grands, alors on peut écrire que

$$S_X \approx \mathrm{N}\Big(\sigma_X, \tfrac{\sigma_X^2}{2m}\Big) \quad \text{et} \quad S_Y \approx \mathrm{N}\Big(\sigma_Y, \tfrac{\sigma_Y^2}{2n}\Big) \tag{6.97}$$

de sorte que

$$Z_0 := \frac{S_X - S_Y}{S_p \left(\frac{1}{2m} + \frac{1}{2n}\right)^{1/2}} \overset{H_0}{\approx} \mathrm{N}(0,1) \tag{6.98}$$

où S_p^2 est défini en (6.85). On rejette alors H_0: $\sigma_X^2 = \sigma_Y^2$ au seuil de signification α si et seulement si

$$\begin{cases} |Z_0| > z_{\alpha/2} \text{ si } H_1\text{: } \sigma_X^2 \neq \sigma_Y^2 \\ Z_0 > z_\alpha \quad \text{ si } H_1\text{: } \sigma_X^2 > \sigma_Y^2 \end{cases} \tag{6.99}$$

Exemple 6.4.10. On considère deux procédés de fabrication pour des bouteilles. Soit $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ la capacité des bouteilles fabriquées selon le procédé utilisé actuellement, et soit $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$ la capacité des bouteilles fabriquées selon un nouveau procédé. Tous les paramètres sont inconnus, mais on croit que $\mu_X = \mu_Y$. On veut tester au seuil de signification $\alpha = 0{,}05$

$$H_0\text{: } \sigma_X^2 = \sigma_Y^2 \quad \text{contre} \quad H_1\text{: } \sigma_X^2 > \sigma_Y^2$$

On prend au hasard 20 bouteilles fabriquées selon le procédé actuel et 25 selon le nouveau procédé. On trouve que $s_X^2 \simeq 0{,}0144$ et $s_Y^2 \simeq 0{,}0064$. On calcule

$$f_0 := \frac{s_X^2}{s_Y^2} \simeq \frac{0{,}0144}{0{,}0064} = 2{,}25$$

Maintenant, on trouve dans une table statistique que $F_{0{,}05;19,24} \simeq 2{,}04$ ($\simeq 2{,}02$ selon la formule d'approximation (5.96) du chapitre 5, page 243). Par conséquent, on peut rejeter H_0 au seuil $\alpha = 0{,}05$.

Remarque. Dans cet exemple, on suppose que les moyennes théoriques μ_X et μ_Y sont inconnues mais égales. Cela ne change rien au test d'hypothèses que l'on utilise. Cette information est donnée pour expliquer que l'on s'intéresse aux variances des variables aléatoires. Cependant, si l'on nous avait donné les valeurs numériques de μ_X et μ_Y, alors on aurait dû en tenir compte dans le test d'hypothèses. On a vu, dans le cas du test d'une variance théorique σ^2, que l'on ne se sert pas de la même statistique lorsque μ est connu. Ici, le test serait basé sur le fait que

$$F_{00}^2 := \frac{\hat{\sigma}_X^2}{\hat{\sigma}_Y^2} \overset{H_0}{\sim} F_{m,n} \tag{6.100}$$

où $\hat{\sigma}_X^2 := \frac{1}{m}\sum_{i=1}^m (X_i - \mu_X)^2$ et $\hat{\sigma}_Y^2 := \frac{1}{n}\sum_{i=1}^n (Y_i - \mu_Y)^2$. ♢

6.4.9 Test de l'égalité de deux proportions

On considère deux populations indépendantes; on veut tester

$$H_0\colon p_1 = p_2 \tag{6.101}$$

où p_i désigne la proportion des individus dans la *population* i ($i = 1, 2$) qui possèdent une certaine caractéristique. Pour ce faire, on prélève un échantillon aléatoire de chacune des deux populations. Soit X_i le nombre d'individus dans l'*échantillon* aléatoire de la population i qui possèdent la caractéristique considérée; alors $X_i \sim \mathrm{B}(n_i, p_i)$, où n_i est la taille de l'échantillon aléatoire de la population i. Si n_i est suffisamment grand, alors on a:

$$X_i \approx \mathrm{N}(n_i p_i, n_i p_i (1 - p_i)) \tag{6.102}$$

Cela implique que

$$\hat{p}_i := \frac{X_i}{n_i} \approx \mathrm{N}\left(p_i, \frac{p_i(1-p_i)}{n_i}\right) \quad (i = 1, 2) \tag{6.103}$$

de sorte que

$$Z := \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx \mathrm{N}(0, 1) \tag{6.104}$$

lorsque H_0 est vraie et $p_1 = p_2 := p$. Puisque p est un paramètre inconnu, il faut l'estimer; on pose:

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} \tag{6.105}$$

La statistique utilisée pour tester H_0 est

$$Z_0 := \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{6.106}$$

On rejette l'hypothèse nulle H_0 au seuil de signification α si et seulement si

$$\begin{cases} |Z_0| > z_{\alpha/2} & \text{si } H_1\colon p_1 \neq p_2 \\ Z_0 > z_\alpha & \text{si } H_1\colon p_1 > p_2 \end{cases} \tag{6.107}$$

Remarques.
(i) Il existe une procédure, basée sur la distribution hypergéométrique, pour tester H_0 lorsque les échantillons aléatoires sont de petites tailles.

(ii) Contrairement au cas du test d'une proportion théorique p, les formules pour la valeur de β et la taille des échantillons sont assez compliquées et ne seront pas données ici. Notons que pour calculer β lorsque l'hypothèse H_1: $p_1 \neq p_2$ est vraie, par exemple, il faut non seulement préciser la valeur de la vraie différence entre p_1 et p_2, mais il faut aussi donner les proportions p_1 et p_2.

(iii) On pourrait généraliser la procédure pour tester H_0: $p_1 - p_2 = \Delta$.

Exemple 6.4.11. On désire tester l'hypothèse qu'un certain candidat dans une élection est aussi populaire chez les électeurs francophones que chez les anglophones. On pose:

$$H_0\colon p_1 = p_2 \quad \text{contre} \quad H_1\colon p_1 \neq p_2$$

où p_1 (respectivement p_2) est la proportion des électeurs francophones (respectivement anglophones) qui appuient le candidat en question. On interroge 200 électeurs francophones et 100 électeurs anglophones pris au hasard; on obtient $x_1 = 76$ et $x_2 = 52$. Ensuite, on calcule

$$\hat{p} = \frac{76 + 52}{200 + 100} \simeq 0{,}4267$$

et

$$z_0 = \frac{\frac{76}{200} - \frac{52}{100}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{200} + \frac{1}{100}\right)}} \simeq -2{,}31$$

Si l'on choisit $\alpha = 0{,}05$, alors on compare cette quantité (en valeur absolue) à 1,960. Étant donné que $|-2{,}31| > 1{,}96$, on peut rejeter l'hypothèse nulle H_0 au seuil $\alpha = 0{,}05$. ♢

(III) Tests de l'égalité de plusieurs paramètres

6.4.10 Test de l'égalité de plusieurs proportions

On considère k populations indépendantes; soit p_j, pour $j = 1, \ldots, k$, la proportion des individus dans la *population* j qui possèdent une certaine caractéristique. On veut faire le test de

$$H_0\colon p_1 = p_2 = \ldots = p_k \quad \text{contre} \quad H_1\colon p_i \neq p_j \tag{6.108}$$

pour au moins un couple (i, j), où $i \neq j$.

Soit X_j le nombre d'individus qui possèdent la caractéristique considérée dans un *échantillon aléatoire* de taille n_j de la population j; alors on a: $X_j \sim \mathrm{B}(n_j, p_j)$ pour $j = 1, \ldots, k$. Si H_0 est vraie, alors la statistique

$$D^2 := \sum_{j=1}^{k} \frac{(X_j - n_j\hat{p})^2}{n_j\hat{p}(1-\hat{p})} \tag{6.109}$$

où

$$\hat{p} := \frac{\sum_{j=1}^{k} X_j}{\sum_{j=1}^{k} n_j} \tag{6.110}$$

présente approximativement une distribution du khi-deux à $k-1$ degrés de liberté. On rejette donc l'hypothèse nulle H_0 si et seulement si

$$D^2 > \chi^2_{\alpha,k-1} \tag{6.111}$$

Remarques.
(i) Il est préférable que la condition $n_j\hat{p} \geq 5$ soit respectée pour tous les j.
(ii) La procédure utilisée est la même que dans le cas du test d'indépendance (voir la page 297).

Exemple 6.4.12. Soit p_j le pourcentage d'étudiants qui travaillent à temps partiel parmi ceux des disciplines suivantes: arts, sciences, agriculture et éducation. On a interrogé 50 étudiants de chacune des disciplines et on a construit le tableau suivant:

Discipline	Arts	Sciences	Agriculture	Éducation	Total
Travail	30	25	15	20	90
Pas de travail	20	25	35	30	110
n_j	50	50	50	50	200

On a: $\hat{p} = 90/200 = 0{,}45$ et $n_j\hat{p} = 50(0{,}45) = 22{,}5 > 5\ \forall j$. Alors on peut calculer

$$d^2 = \frac{(30-22{,}5)^2}{12{,}375} + \frac{(25-22{,}5)^2}{12{,}375} + \frac{(15-22{,}5)^2}{12{,}375} + \frac{(20-22{,}5)^2}{12{,}375} \simeq 10{,}10$$

Puisque $\chi^2_{0{,}05;3} \overset{\text{tab. } 5.3}{\simeq} 7{,}85$, on rejette H_0 au seuil de signification $\alpha = 0{,}05$. $\diamondsuit$

6.4.11 Test de l'égalité de plusieurs moyennes; analyse de la variance

Soit $X_1, \ldots, X_k$ des variables aléatoires *indépendantes*. On considère le modèle

$$X_i \sim \mathrm{N}(\mu + d_i, \sigma^2) \quad \text{pour } i = 1, \ldots, k \tag{6.112}$$

où l'on peut supposer, sans perte de généralité (puisque μ est un paramètre inconnu), que

$$\sum_{i=1}^{k} d_i = 0 \tag{6.113}$$

Remarque. On peut interpréter la quantité d_i comme étant la différence entre la moyenne de X_i et la moyenne μ des k variables aléatoires X_1, …, X_k.

On recueille un échantillon aléatoire de taille n_i de la population i, pour $i = 1, \ldots, k$, et on désire tester l'égalité des k moyennes; c'est-à-dire que l'on pose:

$$H_0\colon d_1 = d_2 = \ldots = d_k = 0 \quad \text{contre} \quad H_1\colon d_i \neq 0 \tag{6.114}$$

pour au moins un i.

Remarque. En pratique, il est préférable que tous les échantillons aléatoires soient de même taille. En effet, le fait que les variances des variables aléatoires ne sont peut-être pas toutes égales en réalité (contrairement à l'hypothèse faite dans le modèle ci-dessus) est alors moins grave; de plus, si l'on recueille km observations en tout, alors la puissance du test est maximisée en choisissant $n_i = m$ pour $i = 1, \ldots, k$.

Soit X_{ij} la j^{e} observation de la variable aléatoire X_i. On définit maintenant les quantités suivantes:

$$\begin{aligned} X_{i\cdot} &= \sum_{j=1}^{n_i} X_{ij}, \qquad \bar{X}_{i\cdot} = \frac{X_{i\cdot}}{n_i} \quad (i = 1, \ldots, k) \\ X_{\cdot\cdot} &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}, \ \bar{X}_{\cdot\cdot} = \frac{X_{\cdot\cdot}}{n}, \quad \text{où } n := \textstyle\sum_{i=1}^{k} n_i \end{aligned} \tag{6.115}$$

Ensuite, on considère la somme des carrés (SC)

$$SC_T := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})\right]^2$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(X_{ij} - \bar{X}_{i.})^2 + (\bar{X}_{i.} - \bar{X}_{..})^2 + 2(X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..})\right]$$
$$= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 \tag{6.116}$$

On écrit:

$$SC_T = SC_I + SC_E \tag{6.117}$$

où SC_I désigne la somme des carrés à l'**intérieur** des échantillons (la source de variation appelée **intra** échantillons) et SC_E la somme des carrés **entre** les échantillons (la source de variation **inter** échantillons). On peut montrer que

$$\frac{SC_T}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{SC_E}{\sigma^2} \overset{H_0}{\sim} \chi^2_{k-1} \quad \text{et} \quad \frac{SC_I}{\sigma^2} \sim \chi^2_{n-k} \tag{6.118}$$

De plus, en utilisant un théorème dû à Cochran, on trouve que SC_I et SC_E sont des variables aléatoires *indépendantes*. Il s'ensuit que

$$F_0 := \frac{SC_E/(k-1)}{SC_I/(n-k)} \overset{H_0}{\sim} F_{k-1,n-k} \tag{6.119}$$

Pour estimer le paramètre inconnu σ^2, on utilise

$$\hat{\sigma}^2 := \frac{SC_I}{n-k} \tag{6.120}$$

On peut montrer que

$$E\left[\hat{\sigma}^2\right] = \sigma^2 \tag{6.121}$$

De même, on trouve que

$$E\left[\frac{SC_E}{k-1}\right] = \sigma^2 + \sum_{i=1}^{k}\frac{n_i d_i^2}{k-1} \tag{6.122}$$

Ainsi, si la contre-hypothèse H_1 est vraie, alors l'espérance mathématique du numérateur dans la définition de F_0 ci-dessus est supérieure à l'espérance mathématique du dénominateur. Par conséquent, le test consiste à rejeter H_0 au seuil de signification α si et seulement si $F_0 > F_{\alpha,k-1,n-k}$.

Pour simplifier les calculs, on utilise les formules suivantes:

$$SC_T := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 - \frac{X_{..}^2}{n} \tag{6.123}$$

et

$$SC_E := \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^{k} \frac{X_{i.}^2}{n_i} - \frac{X_{..}^2}{n} \tag{6.124}$$

La somme des carrés SC_I est alors obtenue en calculant $SC_T - SC_E$.

Finalement, on peut résumer la procédure en un tableau d'**analyse de la variance**:

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens	F_0
Inter	SC_E	$k-1$	$\frac{SC_E}{k-1}$	$\frac{SC_E}{(k-1)\hat{\sigma}^2}$
Intra	SC_I	$n-k$	$\frac{SC_I}{n-k} = \hat{\sigma}^2$	
Totale	SC_T	$n-1$		

Exemple 6.4.13. Soit X_1, X_2 et X_3 des variables aléatoires qui désignent la durée de vie (en semaines) de trois types d'ampoules. On a recueilli les observations x_{ij} suivantes:

X_1	8 6 10
X_2	2 4 5 1
X_3	9 7 7 8 5 6

On suppose que les X_i sont des variables aléatoires indépendantes telles que

$$X_i \sim \mathrm{N}(\mu + d_i, \sigma^2) \quad \text{pour } i = 1, 2, 3$$

et on désire tester

$$H_0\text{: } d_1 = d_2 = d_3 = 0 \quad \text{contre} \quad H_1\text{: } d_i \neq 0$$

pour au moins un i. On calcule

$$x_{1.} = 8 + 6 + 10 = 24, \quad x_{2.} = 12 \quad \text{et} \quad x_{3.} = 42$$

De même,

$$x_{\cdot\cdot} = 24 + 12 + 42 = 78 \quad \text{et} \quad \sum_{i=1}^{3}\sum_{j=1}^{n_i} x_{ij}^2 = 550$$

Ainsi,

$$SC_T = \sum_{i=1}^{3}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{x_{\cdot\cdot}^2}{13} = 550 - \frac{(78)^2}{13} = 82$$

et

$$SC_E = \sum_{i=1}^{3} \frac{x_{i\cdot}^2}{n_i} - \frac{x_{\cdot\cdot}^2}{13} = \frac{(24)^2}{3} + \frac{(12)^2}{4} + \frac{(42)^2}{6} - \frac{(78)^2}{13} = 54$$

On peut maintenant construire le tableau d'analyse de la variance suivant:

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens	F_0
Inter	54	2	27	$\simeq 9{,}64$
Intra	28	10	2,8	
Totale	82	12		

Finalement, on trouve dans le tableau des valeurs de $F_{0,05;n_1,n_2}$ en appendice, à la page 519, que $F_{0,05;2,10} \simeq 4{,}10$. Puisque $9{,}64 > 4{,}10$, on peut rejeter l'hypothèse H_0 d'égalité des moyennes au seuil de signification $\alpha = 0{,}05$. $\diamondsuit$

On peut estimer les paramètres μ et d_i par la *méthode des moindres carrés*. C'est-à-dire que l'on minimise la fonction

$$SC(\mu, d_i) := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \mu - d_i)^2 \tag{6.125}$$

Puisqu'on a posé que $\sum_{i=1}^{k} d_i = 0$, il est naturel d'imposer la condition

$$\sum_{i=1}^{k} n_i \hat{d}_i = 0 \tag{6.126}$$

On trouve alors que

$$\hat{\mu} = \bar{X}_{\cdot\cdot} \quad \text{et} \quad \hat{d}_i = \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot} \tag{6.127}$$

De là, on déduit que $\hat{\mu}_i = \bar{X}_{i.}$, où $\mu_i := \mu + d_i$ est la moyenne de la variable aléatoire X_i.

Maintenant, on peut écrire que

$$\bar{X}_{i.} \sim \mathrm{N}\Big(\mu_i, \tfrac{\sigma^2}{n_i}\Big) \quad \text{et} \quad \bar{X}_{..} \sim \mathrm{N}\Big(\mu, \tfrac{\sigma^2}{n}\Big) \tag{6.128}$$

Par conséquent, on peut montrer que des intervalles de confiance bilatéraux à $100(1-\alpha)$ % sont donnés par:

$$\text{pour } \mu\text{: } \bar{X}_{..} \pm t_{\alpha/2,n-k}\frac{\hat{\sigma}}{\sqrt{n}}, \tag{6.129}$$

$$\text{pour } \mu_i\text{: } \bar{X}_{i.} \pm t_{\alpha/2,n-k}\frac{\hat{\sigma}}{\sqrt{n_i}} \tag{6.130}$$

$$\text{pour } \mu_i - \mu_j = d_i - d_j\text{: } \bar{X}_{i.} - \bar{X}_{j.} \pm t_{\alpha/2,n-k}\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \tag{6.131}$$

où $\hat{\sigma} = \sqrt{SC_I/(n-k)}$.

Exemple 6.4.14. Un intervalle de confiance à 95 % pour μ dans l'exemple précédent est le suivant:

$$\frac{78}{13} \pm t_{0,025;13-3}\frac{\sqrt{2{,}8}}{\sqrt{13}} \overset{\text{tab. } 5.2}{\simeq} 6 \pm 1{,}03$$

♢

6.5 Exercices du chapitre 6

Exercices résolus

Question n° 1

On suppose que le temps de service (en minutes) à un comptoir est une variable aléatoire X dont la fonction de densité est donnée par

$$f_X(x;\theta) = 2\theta^2 x^2 e^{-\theta x^2} \quad \text{pour } x \geq 0$$

où $\theta > 0$. On a recueilli les observations suivantes de X:

Intervalle	$[0,1)$	$[1,2)$	$[2,3)$	$[3,\infty)$
Effectif	10	55	30	5

Tester, au seuil de signification $\alpha = 0{,}05$, l'ajustement du modèle considéré aux données, avec $\theta = 1/2$.

Indication. On a: $P[X \leq x] = 1 - \left(1 + \theta x^2\right) e^{-\theta x^2}$ pour $x \geq 0$.

Solution. On veut tester

$$H_0 \colon f_X(x) = \frac{1}{2} x^2 e^{-x^2/2}$$

au seuil de signification $\alpha = 0{,}05$. On a:

$$P[X \leq x] = 1 - \left(1 + \tfrac{1}{2}x^2\right) e^{-x^2/2} \quad \text{pour } x \geq 0$$

De là, on obtient le tableau suivant:

Intervalle	$[0,1)$	$[1,2)$	$[2,3)$	$[3,\infty)$	$\sum$
n_i	10	55	30	5	100
$(\simeq)\ p_i$	0,090	0,504	0,345	0,061	1
$100p_i = m_i$	9	50,4	34,5	6,1	100

On calcule (puisque $m_i > 5\ \forall i$)

$$d^2 = \frac{(10-9)^2}{9} + \frac{(55-50{,}4)^2}{50{,}4} + \frac{(30-34{,}5)^2}{34{,}5} + \frac{(5-6{,}1)^2}{6{,}1} \simeq 1{,}32$$

Maintenant, on a: $D^2 \overset{H_0}{\approx} \chi^2_{4-0-1}$ et $\chi^2_{0{,}05;3} \overset{\text{tab. } 5.3}{\simeq} 7{,}815$. Finalement, puisque $1{,}32 < 7{,}815$, on ne peut pas rejeter H_0 au seuil $\alpha = 0{,}05$. Donc, on accepte le modèle proposé avec $\theta = 1/2$.

Question n$^{\text{o}}$ 2

Dans un sondage effectué auprès de 1000 personnes, 645 d'entre elles ont dit préférer l'option A aux autres options dans une question. Soit p la popularité de l'option A.

(a) D'après les données recueillies, peut-on conclure que la valeur de p est significativement supérieure à 60 %? Utiliser $\alpha = 0{,}05$.

(b) Quelle est, approximativement, la plus petite valeur de α pour laquelle on rejette l'hypothèse nulle en (a)?

Solution. (a) On accepte H_1: $p > 0{,}60$ si et seulement si

$$\hat{p} > 0{,}60 + \underbrace{z_{0{,}05}}_{1{,}645} \sqrt{\frac{0{,}60 \times 0{,}40}{1000}} \simeq 0{,}6255$$

Étant donné que $\hat{p} = \frac{645}{1000} = 0{,}645 > 0{,}6255$, on peut conclure que p est significativement supérieur à 0,60 si l'on prend $\alpha = 0{,}05$.

(b) On cherche la valeur de α pour laquelle

$$0{,}645 = 0{,}60 + z_\alpha \sqrt{\frac{0{,}60 \times 0{,}40}{1000}} \iff z_\alpha \simeq 2{,}90$$

Puisque $P[\mathrm{N}(0,1) \leq 2{,}90] \overset{\text{tab. A.3}}{\simeq} 0{,}998$, on conclut que $\alpha_{\min} \simeq 0{,}002$.

Question n$^{\text{o}}$ 3

Un échantillon aléatoire particulier de neuf cigarettes d'une certaine marque contient, en moyenne, 4,2 mg de nicotine par cigarette. Le fabricant affirme que la teneur X en nicotine de ses cigarettes ne dépasse pas, en moyenne, 3,5 mg. De plus, on suppose que $X \approx \mathrm{N}(\mu; 1{,}96)$.

(a) Peut-on mettre en doute les affirmations du fabricant? Utiliser $\alpha = 0{,}05$.

(b) Quelle est la probabilité de conclure, en (a), que le fabricant a tort si $\mu = 3{,}8$?

(c) Quel est le nombre minimal de cigarettes que l'on devrait examiner si l'on veut que la probabilité calculée en (b) soit supérieure ou égale à 90 %?

(d) Pour mieux protéger le consommateur, est-il préférable d'effectuer le test en (a) avec $\alpha = 0{,}05$ ou $\alpha = 0{,}01$? Justifier votre réponse.

Solution. (a) On veut tester

$$H_0\colon \mu = 3{,}5 \quad \text{contre} \quad H_1\colon \mu > 3{,}5$$

au seuil $\alpha = 0{,}05$. On calcule

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{4{,}2 - 3{,}5}{1{,}4/\sqrt{9}} = 1{,}5$$

On rejette H_0 si et seulement si $z_0 > z_{0{,}05} \overset{\text{tab. 5.1}}{\simeq} 1{,}645$. Donc, on ne peut pas rejeter H_0 au seuil $\alpha = 0{,}05$.

(b) On cherche $1-\beta\,(\mu=3{,}8)$. Or, on a:

$$\begin{aligned}\beta\,(\mu=3{,}8) &= \Phi\left(z_{0,05}-\frac{\mu-\mu_0}{\sigma}\sqrt{n}\right)\simeq\Phi\left(1{,}645-\frac{3{,}8-3{,}5}{1{,}4}\sqrt{9}\right)\\ &\simeq\Phi\,(1{,}00)\simeq 0{,}8413\end{aligned}$$

de sorte que

$$1-\beta\,(\mu=3{,}8)\simeq 0{,}1587$$

(c) On veut que

$$1-\beta\,(\mu=3{,}8)\geq 0{,}90 \iff \beta\,(\mu=3{,}8)\leq 0{,}10$$

Il faut prendre

$$n\geq\frac{(z_{0,05}+z_{0,10})^2\,(1{,}4)^2}{(3{,}8-3{,}5)^2}\overset{\text{tab. 5.1}}{\simeq}(1{,}645+1{,}282)^2\,\frac{(1{,}4)^2}{(0{,}3)^2}\simeq 186{,}6$$

Donc, il faudrait examiner au moins 187 cigarettes.

(d) Pour mieux protéger le consommateur, il faut choisir le plus grand α, car plus α est petit, plus il est difficile de conclure que le fabricant a tort. Donc, il est préférable de choisir $\alpha=0{,}05$.

Question n$^{\text{o}}$ 4

On veut comparer la force de compression (en kilogrammes par centimètre carré) de deux types de béton. On veut montrer qu'en moyenne le second est plus fort que le premier. Des mesures faites sur deux échantillons aléatoires particuliers sont données ci-dessous:

Béton 1	295 319 304 302
Béton 2	318 316 312 318

(a) Définir les variables et spécifier les hypothèses de base nécessaires.

(b) Comparer, au seuil de signification $\alpha=0{,}05$, les forces de compression moyennes en supposant que $\sigma_1^2\neq\sigma_2^2$.

(c) Montrer, par un test d'hypothèses, qu'en fait les variances ne sont pas significativement différentes au seuil de signification $\alpha=0{,}05$.

(d) Refaire le test de comparaison des moyennes de la partie (b) avec $\sigma_1^2=\sigma_2^2$.

Solution. On a:

$$n_1 = 4,\ \bar{x}_1 = 305,\ s_1^2 = 102$$
$$n_2 = 4,\ \bar{x}_2 = 316,\ s_2^2 = 8$$

(a) On pose: X_i = force de compression du béton i, pour $i = 1, 2$. On suppose que $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ et que X_1 et X_2 sont des variables aléatoires indépendantes.
(b) On veut tester

$$H_0\colon \mu_1 = \mu_2 \quad \text{contre} \quad H_1\colon \mu_1 < \mu_2$$

au seuil de signification $\alpha = 0{,}05$. On calcule

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{305 - 316}{\sqrt{\frac{102}{4} + \frac{8}{4}}} \simeq -2{,}10$$

On rejette H_0 si et seulement si $t_0 < -t_{0{,}05;\nu}$, où

$$\nu = \frac{(25{,}5 - 2)^2}{\frac{(25{,}5)^2}{3} + \frac{2^2}{3}} \simeq 3{,}47$$

Maintenant, on a:

$$-t_{0{,}05;3{,}47} < -t_{0{,}05;4} \overset{\text{tab. } 5.2}{\simeq} -2{,}132$$

et t_0 n'est pas inférieur à $-2{,}132$. Alors on conclut que l'on ne peut pas rejeter H_0 au seuil $\alpha = 0{,}05$.
(c) On veut tester

$$H_0\colon \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1\colon \sigma_1^2 \neq \sigma_2^2$$

au seuil de signification $\alpha = 0{,}05$. On calcule

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{102}{8} = 12{,}75$$

On rejette H_0 au seuil $\alpha = 0{,}05$ si et seulement si

$$f_0 > F_{0{,}025;3,3} \overset{\text{tab. } 5.4}{\simeq} 15{,}4 \quad \text{ou} \quad f_0 < F_{0{,}975;3,3} \simeq \frac{1}{15{,}4} \simeq 0{,}065$$

Donc, on doit conclure que les variances ne sont pas significativement différentes au seuil $\alpha = 0{,}05$.

(d) Puisque l'on peut supposer que $\sigma_1^2 = \sigma_2^2$, on calcule

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où

$$s_p^2 = \frac{3s_1^2 + 3s_2^2}{4+4-2} = \frac{102+8}{2} = 55$$

On trouve que $t_0 \simeq -2{,}10$ et, puisque

$$t_0 < -t_{0{,}05;4+4-2} \overset{\text{tab. } 5.2}{\simeq} -1{,}943$$

on peut rejeter H_0 au seuil de signification $\alpha = 0{,}05$.

Question n° 5

Soit $f_X(x;\theta) = 1/\theta$ pour $0 < x < \theta$. On veut tester

$$H_0 \colon \theta = 1 \quad \text{contre} \quad H_1 \colon \theta > 1$$

Pour ce faire, on recueille *une* observation X_1 de X et on rejette H_0 si et seulement si $X_1 > 0{,}9$.

(a) Quel est le risque de première espèce α du test?

(b) Quelle est la valeur de β si $\theta = 1{,}5$?

Solution. (a) On cherche

$$\alpha = P\left[X_1 > 0{,}9 \mid X_1 \sim \mathrm{U}\,(0,1)\right] = 1 - 0{,}9 = 0{,}1$$

(b) On a:

$$\beta = P\left[X_1 \leq 0{,}9 \mid X_1 \sim \mathrm{U}\,(0;1{,}5)\right] = \frac{0{,}9}{1{,}5} = 0{,}6$$

Question n° 6

Soit $X \sim \mathrm{N}\big(\mu, \sigma^2\big)$. Un échantillon aléatoire particulier de taille $n = 25$ de X a donné $\bar{x} = 0{,}2$ et $s = 0{,}5$.

(a) On veut tester $H_0 \colon \mu = 0$ contre $H_1 \colon \mu \neq 0$ au seuil de signification $\alpha = 0{,}05$. Calculer la valeur de la statistique utilisée pour effectuer le test.

(b) On désire aussi tester $H_0 \colon \sigma^2 = 0{,}2$ contre $H_A \colon \sigma^2 > 0{,}2$ au seuil de signification $\alpha = 0{,}05$. Quelle est la valeur du centile utilisé pour prendre la décision (c'est-à-dire la valeur critique)?

Solution. (a) On rejette H_0 si et seulement si

$$|t_0| = \left| \frac{\bar{x} - 0}{s/\sqrt{n}} \right| = \left| \frac{0{,}2 - 0}{0{,}5/\sqrt{25}} \right| = |2| > t_{0{,}025;24}$$

Donc, la valeur de la statistique utilisée est 2.

(b) On rejette H_0 si et seulement si

$$w_0^2 = \frac{24s^2}{0{,}2} > \chi^2_{0{,}05;24}$$

En utilisant l'approximation de Wilson-Hilferty, on trouve que

$$\chi^2_{0{,}05;24} \simeq 24 \left[1{,}645 \left(\frac{2}{9 \times 24} \right)^{1/2} + 1 - \frac{2}{9 \times 24} \right]^3 \simeq 36{,}41$$

L'approximation de Fisher donne

$$\chi^2_{0{,}05;24} \simeq \frac{1}{2} \left[1{,}645 + (48 - 1)^{1/2} \right]^2 \simeq 36{,}13$$

En fait, on trouve, dans une table statistique ou en utilisant un logiciel, que $\chi^2_{0{,}05;24} \simeq 36{,}415$. Donc, ici l'approximation de Wilson-Hilferty est supérieure à celle de Fisher.

Finalement, si l'on effectue une interpolation linéaire entre les valeurs $\chi^2_{0{,}05;20} \overset{\text{tab. 5.3}}{\simeq} 31{,}41$ et $\chi^2_{0{,}05;25} \simeq 37{,}65$, alors on obtient que $\chi^2_{0{,}05;24} \simeq 36{,}40$, ce qui est une très bonne approximation.

Question n^o 7

On croit que le service de courrier A est plus rapide que le service B. Des observations indépendantes du temps de livraison (en jours) des deux services de courrier ont donné:

$$n_A = 5, \quad \bar{x}_A = 1,2; \quad n_B = 10, \quad \bar{x}_B = 1{,}5$$

De plus, on suppose que les écarts-types des temps de livraison sont $\sigma_A = \sigma_B = 0{,}2$.

(a) Quelles hypothèses additionnelles doit-on formuler pour pouvoir effectuer la comparaison? Poser aussi les hypothèses H_0 et H_1.

(b) Quel est le plus petit α pour lequel on peut rejeter H_0 si $z_0 = -2{,}74$?

Solution. (a) Soit X_A (respectivement X_B) le temps de livraison avec le service A (respectivement B). On suppose que X_A et X_B présentent une distribution normale, de sorte que

$$X_A \sim \mathrm{N}\left(\mu_A, \sigma_A^2 = 0{,}04\right) \quad \text{et} \quad X_B \sim \mathrm{N}\left(\mu_B, \sigma_B^2 = 0{,}04\right)$$

De plus, on suppose que X_A et X_B sont deux variables aléatoires indépendantes. Enfin, on pose:

$$H_0 \colon \mu_A = \mu_B \quad \text{et} \quad H_1 \colon \mu_A < \mu_B$$

(b) On rejette H_0 si et seulement si

$$z_0 < -z_\alpha \iff -2{,}74 < -z_\alpha$$

Or, on a:

$$\Phi\left(2{,}74\right) \overset{\text{tab. A.3}}{\simeq} 0{,}9969$$

Alors on peut écrire que

$$\alpha_{\min} \simeq (1 - 0{,}9969)^+ = 0{,}0031^+$$

Question n$^{\text{o}}$ 8

On s'intéresse à la durée de vie X (en dizaines de milliers de kilomètres) des pneus de marque A. On désire tester $H_0 \colon \mu_X = 50$ contre $H_1 \colon \mu_X < 50$ au seuil de signification $\alpha = 0{,}05$. On suppose que $X \sim \mathrm{N}(\mu_X, 25)$.

(a) Quelle est la valeur de β si $n = 9$ et $\mu_X = 45$?

(b) Quelle est la plus petite valeur de n pour laquelle $\beta \leq 0{,}10$ si $\mu_X = 45$?

Solution. (a) On a:

$$\beta\left(\mu_X = 45\right) = \Phi\left(z_{0{,}05} + \frac{(45 - 50)}{5}\sqrt{9}\right) \simeq \Phi\left(-1{,}355\right) \overset{\text{tab. A.3}}{\simeq} 0{,}088$$

(b) On calcule

$$n = \frac{\left(z_{0{,}05} + z_{0{,}10}\right)^2 \, 25}{\left(45 - 50\right)^2} \overset{\text{tab. 5.1}}{\simeq} (1{,}645 + 1{,}282)^2 \simeq 8{,}6 \implies n_{\min} = 9$$

Question n° 9

Soit $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ deux variables aléatoires indépendantes. Des échantillons aléatoires particuliers de X_1 et X_2, respectivement, ont donné:

$$n_1 = 9, \quad \bar{x}_1 = 5, \quad s_1 = 2; \quad n_2 = 10, \quad \bar{x}_2 = 3, \quad s_2 = 1$$

(a) La variance σ_1^2 est-elle significativement supérieure à σ_2^2? Utiliser $\alpha = 0{,}05$.

(b) Si l'on suppose que les variances théoriques sont égales, peut-on alors dire que les moyennes théoriques sont significativement différentes? Utiliser $\alpha = 0{,}05$.

Solution. (a) On accepte H_1: $\sigma_1^2 > \sigma_2^2$ si et seulement si

$$f_0 = \frac{2^2}{1^2} = 4 > F_{0{,}05;9-1,10-1} \overset{\text{p. 519}}{\simeq} 3{,}23$$

Donc, on accepte H_1 au seuil de signification $\alpha = 0{,}05$.

(b) On accepte H_1: $\mu_1 \neq \mu_2$ si et seulement si

$$|t_0| \simeq \left| \frac{5-3}{s_p \sqrt{\frac{1}{9} + \frac{1}{10}}} \right| \simeq 2{,}80 > t_{0{,}025;9+10-2}$$

car

$$s_p^2 = \frac{8 \cdot 4 + 9 \cdot 1}{9 + 10 - 2} \simeq 2{,}41$$

Or, on a:

$$t_{0{,}025;17} < t_{0{,}025;15} \overset{\text{tab. 5.2}}{\simeq} 2{,}131$$

et $2{,}80 > 2{,}131$. Donc, on accepte H_1 au seuil de signification $\alpha = 0{,}05$.

Question n° 10

Lors d'une étude, on a trouvé que 5 des 100 appareils de marque A vendus pendant une certaine période avaient dû être réparés au cours de leur période de garantie. Soit θ la proportion véritable des appareils de cette marque devant être réparés avant la fin de leur période de garantie.

(a) Combien d'appareils auraient dû être réparés pour que l'on puisse conclure que $\theta > 0{,}03$, avec $\alpha = 0{,}05$?

(b) Si $\theta = 0{,}03$, quelle distribution présente approximativement le nombre X d'appareils, parmi les 100 vendus, qui doivent être réparés pendant leur période de garantie? Donner aussi le ou les paramètres de cette distribution.

Solution. (a) On accepte H_1: $\theta > 0{,}03$ si et seulement si le nombre d'appareils devant être réparés est supérieur à $100\,C$, où

$$C = 0,03 + z_{0,05}\left[\frac{(0{,}03)\,(0{,}97)}{100}\right]^{1/2} \simeq 0{,}058$$

Alors cela aurait nécessité au moins six appareils.

Remarque. Si l'on avait fait la correction de continuité, alors le nombre minimal d'appareils aurait été de sept.

(b) On a:

$$X \sim \mathrm{B}\,(n = 100, p = 0{,}03) \approx \mathrm{Poi}(3)$$

Remarque. On peut aussi écrire que $X \approx \mathrm{N}(3; 2{,}91)$, par le théorème central limite, mais cette approximation est moins justifiée car $np = 3 < 5$ (voir la page 90).

Question n° 11

On cherche à déterminer si le revenu des parents a une influence sur le succès de leurs enfants à l'école. On dispose des données suivantes:

$Y \setminus X$	Faible	Moyen	Élevé
Faible	10	20	10
Moyen	10	40	10
Bon	10	30	10

où X = revenu des parents et Y = succès scolaire des enfants.

(a) Calculer les effectifs sous l'hypothèse H_0 de l'indépendance entre les variables X et Y.

(b) On trouve que $d^2 \simeq 2{,}8$. Quelle est alors la conclusion du test si $\alpha = 0{,}05$? Donner aussi la valeur du centile utilisé pour prendre la décision.

Solution. (a) On a, sous l'hypothèse d'indépendance entre X et Y:

$Y \setminus X$	Faible	Moyen	Élevé	
Faible	8	24	8	40
Moyen	12	36	12	60
Bon	10	30	10	50
	30	90	30	150

(b) On rejette l'hypothèse H_0 si et seulement si

$$d^2 > \chi^2_{0,05;(3-1)(3-1)} \overset{\text{tab. } 5.3}{\simeq} 9{,}49$$

Puisque $2{,}8 < 9{,}49$, on ne rejette pas H_0 au seuil de signification $\alpha = 0{,}05$.

Question n^o 12

On veut tester l'hypothèse H_0 qu'une variable aléatoire X présente une distribution $\mathrm{B}(3, \frac{1}{2})$. On prélève un échantillon aléatoire particulier de taille 40 de X. On obtient:

i	0	1	2	3
n_i	7	12	18	3

(a) Calculer les effectifs sous H_0, c'est-à-dire les m_i, pour $i = 0, 1, 2$ et 3.

(b) Si $d^2 = 2{,}8$ et $\alpha = 0{,}05$, quelle est la conclusion du test? Donner aussi la valeur du centile utilisé pour effectuer le test.

Solution. (a) On a:

$$p_i = P\left[\mathrm{B}\left(3, \tfrac{1}{2}\right) = i\right] = \binom{3}{i}\left(\tfrac{1}{2}\right)^3$$

pour $i = 0, 1, 2$ et 3. Alors $m_0 = m_3 = 40\,(1/8) = 5$ et $m_1 = m_2 = 15$.

(b) On rejette H_0 au seuil de signification $\alpha = 0{,}05$ si et seulement si

$$d^2 = 2{,}8 > \chi^2_{0,05;4-0-1} \overset{\text{tab. } 5.3}{\simeq} 7{,}81$$

Donc, on accepte H_0.

Question n^o 13

Soit X une variable aléatoire continue. On veut tester H_0: $f_X(x) = 1$, si $0 < x < 1$, au seuil de signification $\alpha = 0{,}10$. Un échantillon aléatoire particulier de taille 50 de X a donné:

Intervalle	$[0, \frac{1}{4})$	$[\frac{1}{4}, \frac{1}{2})$	$[\frac{1}{2}, \frac{3}{4})$	$[\frac{3}{4}, 1]$
Effectif	10	15	12	13

(a) Calculer la statistique utilisée pour effectuer le test.

(b) Si, avec d'autres données, on obtient $d^2 = 10$, quel est alors le plus petit α pour lequel l'hypothèse H_0 est rejetée?

Solution. (a) On a: $m_i = 50 \times \frac{1}{4} = 12{,}5\ (> 5)\ \forall i$. Alors

$$d^2 = \frac{(10-12{,}5)^2}{12{,}5} + \frac{(15-12{,}5)^2}{12{,}5} + \frac{(12-12{,}5)^2}{12{,}5} + \frac{(13-12{,}5)^2}{12{,}5} = 1{,}04$$

(b) On rejette H_0 au seuil de signification α si et seulement si $d^2 = 10 > \chi^2_{\alpha,4-0-1}$. Or, on a: $\chi^2_{0{,}01;3} \overset{\text{tab. 5.3}}{\simeq} 11{,}34$ et $\chi^2_{0{,}025;3} \simeq 9{,}35$. Donc, en effectuant une interpolation linéaire, on peut écrire que

$$\alpha_{\min} \simeq 0{,}02$$

Remarques.
(i) On trouve, avec un logiciel, que

$$P[\chi^2_3 > 10] \simeq 0{,}0186 \quad (\simeq \alpha_{\min})$$

(ii) En se servant de l'approximation de Wilson-Hilferty, on obtient que

$$z_\alpha \simeq 2{,}08 \implies \alpha_{\min} \simeq 0{,}02$$

également.

Question n$^{\text{o}}$ 14

On suppose que tous les étudiant(e)s qui suivent un certain cours possèdent sensiblement les mêmes aptitudes. On suppose aussi que la note d'un(e) étudiant(e) à l'examen final présente (approximativement) une distribution normale. On a les résultats suivants obtenus à cet examen par les étudiant(e)s de deux sections, dont les professeurs sont différents:

Section	n	$\bar{x}$	s
1	49	10,5	4,1
2	35	9,3	3,8

Soit X_i la note qu'un(e) étudiant(e) qui suit le cours dans la section i aura à l'examen final, pour $i = 1, 2$. On suppose que $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ et $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ sont des variables aléatoires indépendantes.

(a) Peut-on conclure que la différence entre les variations dans les notes est significative si l'on utilise un seuil de signification $\alpha = 0{,}10$? Préciser l'hypothèse nulle H_0 et la contre-hypothèse H_1.

Indication. Utiliser le fait que $F_{0{,}05;48,34} > F_{0{,}05;50,50}$.

(b) Peut-on conclure que le professeur a un effet significatif sur les résultats de ses étudiant(e)s? Utiliser $\alpha = 0{,}05$ et préciser H_0 et H_1.

(c) Quel est le nombre minimal d'observations dont on a besoin pour pouvoir détecter, avec une probabilité d'au moins 90 %, une différence réelle d'un point entre les moyennes μ_1 et μ_2?

(d) Quelle est, approximativement, la valeur de β pour le test effectué en (b) si $\mu_1 - \mu_2 = 2$?

Solution. (a) On veut tester

$$H_0\colon \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1\colon \sigma_1^2 \neq \sigma_2^2$$

au seuil de signification $\alpha = 0{,}10$. On calcule

$$f_0 = \frac{s_1^2}{s_2^2} = \left(\frac{4{,}1}{3{,}8}\right)^2 \simeq 1{,}164$$

On rejette H_0 si et seulement si

$$f_0 > F_{0{,}05;49-1,35-1} \quad \text{ou} \quad f_0 < F_{0{,}95;49-1,35-1}$$

Or, on a:

$$F_{0{,}05;48,34} > F_{0{,}05;50,50} \overset{\text{tab. 5.4}}{\simeq} 1{,}60 \quad \text{et} \quad F_{0{,}95;48,34} < 1$$

Donc, on ne rejette pas H_0 au seuil $\alpha = 0{,}10$.

(b) On veut maintenant tester

$$H_0\colon \mu_1 = \mu_2 \quad \text{contre} \quad H_1\colon \mu_1 \neq \mu_2$$

au seuil de signification $\alpha = 0{,}05$. Étant donné que $n_1 > 30$ et $n_2 > 30$, on peut se servir du test où l'on suppose que les écarts-types de X_1 et X_2 sont connus. On rejette H_0 si et seulement si

$$|\bar{x}_1 - \bar{x}_2| > \underbrace{z_{0{,}025}}_{1{,}960} \sqrt{\frac{(4{,}1)^2}{49} + \frac{(3{,}8)^2}{35}} \simeq 1{,}70$$

Puisque $|\bar{x}_1 - \bar{x}_2| = 1{,}2$, on ne rejette pas H_0 au seuil $\alpha = 0{,}05$. C'est-à-dire que le professeur n'a pas d'effet significatif sur les résultats de ses étudiant(e)s.

(c) On doit avoir:

$$n \simeq \frac{(z_{0,025} + z_{0,10})^2 \left[(4,1)^2 + (3,8)^2\right]}{(1)^2} \overset{\text{tab. } 5.1}{\simeq} 328,5$$

Donc cela exige, en tout, environ $2 \times 329 = 658$ (ou 657) observations.

(d) On a (voir la page 314):

$$\delta = \frac{2-0}{\sqrt{\frac{(4,1)^2}{49} + \frac{(3,8)^2}{35}}} \simeq 2,30$$

Alors

$$\begin{aligned}\beta(\delta) &\simeq \Phi(1,960 - 2,30) - \Phi(-1,960 - 2,30) \\ &= \Phi(-0,34) - \Phi(-4,26) \overset{\text{tab. A.3}}{\simeq} (1 - 0,63) - 0 = 0,37\end{aligned}$$

Question n° 15

Un certain manufacturier de caoutchouc synthétique affirme que la dureté moyenne de son caoutchouc est de 64,3 degrés Shore. Des expériences précédentes indiquent que l'écart-type de la dureté est de 2 degrés Shore. On croit que l'affirmation du manufacturier sous-estime ou surestime la dureté moyenne du caoutchouc; alors on effectue un test. Si la vraie moyenne (μ) est de 64,3 degrés, la probabilité d'arriver à cette conclusion doit être de 0,95. De plus, si la différence entre μ et 64,3 est de $\pm$ 1 degré, le test utilisé devrait conclure que μ n'est pas égal à 64,3, et ce, avec une probabilité d'au moins 0,9. On suppose que la dureté du caoutchouc présente une distribution normale.

(a) Combien d'observations doit-on prélever?

(b) Si $\bar{x} = 65$ et n est la valeur calculée en (a), quelle est la conclusion du test?

(c) Calculer la probabilité de conclure que la dureté moyenne du caoutchouc est de 64,3, si la vraie dureté moyenne est de 65 degrés Shore et si l'on utilise la taille de l'échantillon calculée en (a).

(d) Combien d'observations doit-on prélever si l'on veut rejeter H_0 avec un risque de première espèce de 0,05 lorsque $\bar{x} = 65$?

Solution. (a) On a: $\alpha = 0,05$ et $\beta\,(|\mu - 64,3| = 1) = 1 - 0,9 = 0,1$. Alors il faut prendre

$$n \simeq \frac{(z_{0,025} + z_{0,10})^2\,(2)^2}{(\mu - 64,3)^2} \overset{\text{tab. } 5.1}{\simeq} (1,960 + 1,282)^2\,\frac{4}{1} \simeq 42,04$$

Donc, il faut prélever $n = 43$ (ou 42) observations.

(b) On veut tester

$$H_0 : \mu = 64{,}3 \quad \text{contre} \quad H_1 : \mu \neq 64{,}3$$

au seuil de signification $\alpha = 0{,}05$. On rejette H_0 si et seulement si

$$\bar{x} < C_1 = 64{,}3 - z_{0{,}025}\frac{2}{\sqrt{43}} \simeq 64{,}3 - 0{,}6 = 63{,}7$$

ou

$$\bar{x} > C_2 = 64{,}3 + 1{,}960\frac{2}{\sqrt{43}} \simeq 64{,}3 + 0{,}6 = 64{,}9$$

Étant donné que $\bar{x} = 65 > 64{,}9$, on rejette l'hypothèse nulle H_0 au seuil $\alpha = 0{,}05$.

Remarque. Si $\bar{x} = 65$ avec $n = 42$ (plutôt que 43), alors $C_2 \simeq 64{,}9$ aussi et la conclusion est la même.

(c) On veut calculer $\beta\,(\mu = 65)$. On peut écrire que

$$\begin{aligned}
\beta\,(\mu = 65) &\overset{\text{(b)}}{\simeq} P\left[63{,}7 \leq \bar{X} \leq 64{,}9 \mid \mu = 65\right] \\
&= P\left[63{,}7 \leq \mathrm{N}\Big(65, \tfrac{2}{\sqrt{43}}\Big) \leq 64{,}9\right] \\
&\simeq P\left[-4{,}26 \leq \mathrm{N}(0,1) \leq -0{,}33\right] \\
&\simeq P\left[\mathrm{N}(0,1) \leq -0{,}33\right] \overset{\text{tab. A.3}}{\simeq} 0{,}37
\end{aligned}$$

Remarque. On aurait pu se servir directement de la formule (6.41), à la page 302.

(d) On cherche n (minimal) tel que

$$65 > 64{,}3 + z_{0{,}025}\frac{2}{\sqrt{n}} \iff \sqrt{n} > 5{,}6 \iff n > 31{,}3$$

Donc, il faut prélever au moins 32 observations.

Question n$^{\text{o}}$ 16

On a pris 200 articles au hasard parmi ceux produits par une certaine machine. On a trouvé que 25 d'entre eux étaient défectueux. On a alors procédé à un ajustement de la machine. Une semaine plus tard, on a de nouveau pris 200 articles au hasard parmi ceux produits par la machine en question et on a trouvé que 15 d'entre eux étaient défectueux.

(a) Peut-on conclure que la machine fonctionne mieux que la semaine précédente? Formuler le problème sous la forme d'un test d'hypothèses et utiliser un seuil de signification α de 5 %. Quelle hypothèse de base doit-on faire pour pouvoir effectuer le test?

(b) Quelle est la plus petite valeur de α pour laquelle on peut rejeter l'hypothèse nulle en (a)?

(c) Parmi les 200 articles pris au hasard avant l'ajustement de la machine, outre les 25 articles défectueux, 13 ne pouvaient être vendus que comme articles de deuxième qualité. De plus, parmi les 25 défectueux, 2 articles pouvaient être réparés et vendus comme articles de deuxième qualité. Soit X la variable aléatoire qui prend la valeur

0 si un article donné est défectueux et ne peut être réparé;
1 si l'article ne peut être vendu que comme article de deuxième qualité;
2 autrement.

Tester, au seuil de signification $\alpha = 0{,}05$, l'ajustement du modèle

$$p_X(x) = \begin{cases} 0{,}10 \text{ si } x = 0 \\ 0{,}10 \text{ si } x = 1 \\ 0{,}80 \text{ si } x = 2 \end{cases}$$

aux données.

Solution. (a) Soit p_1 (respectivement p_2) la proportion d'articles défectueux produits par la machine avant son ajustement (respectivement une semaine après son ajustement); on veut tester

$$H_0\colon p_1 = p_2 \quad \text{contre} \quad H_1\colon p_1 > p_2$$

au seuil de signification $\alpha = 0{,}05$. On doit supposer que le pourcentage d'articles défectueux avant l'ajustement et le pourcentage une semaine après l'ajustement sont *indépendants*. Le test consiste à rejeter H_0 si et seulement si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{200} + \frac{1}{200}\right)}} > z_{0{,}05}$$

où

$$\hat{p}_1 = \frac{25}{200}, \quad \hat{p}_2 = \frac{15}{200} \quad \text{et} \quad \hat{p} = \frac{25+15}{200+200}$$

C'est-à-dire que l'on rejette H_0 si et seulement si

$$\frac{0{,}05}{0{,}03} \simeq 1{,}67 > z_{0{,}05} \overset{\text{tab. } 5.1}{\simeq} 1{,}645$$

Donc, on peut rejeter H_0 au seuil de signification $\alpha = 0{,}05$ et conclure que la machine fonctionne mieux que la semaine précédente.

(b) On cherche α tel que $1{,}67 \simeq z_\alpha$. Puisque

$$P\left[\mathrm{N}(0,1) \leq 1{,}67\right] \overset{\text{tab. A.3}}{\simeq} 0{,}952$$

on peut écrire que

$$\alpha_{\min} \simeq 0{,}048^{+}$$

(c) On construit le tableau suivant:

i	0	1	2	$\sum$
n_i	23	15	162	$200 = n$
p_i	0,1	0,1	0,8	1
$np_i = m_i$	20	20	160	200

Étant donné que $m_i \geq 5\ \forall i$, on calcule

$$d^2 = \frac{(23-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(162-160)^2}{160} = 1{,}725$$

On rejette le modèle au seuil de signification $\alpha = 0{,}05$ si et seulement si

$$d^2 > \chi^2_{0{,}05;3-1} \overset{\text{tab. } 5.3}{\simeq} 5{,}99$$

Donc, on accepte le modèle si $\alpha = 0{,}05$.

Question n° 17

La résistance ohmique d'un certain composant électronique doit être, en moyenne, de 400 ohms. Un échantillon de 16 composants, prélevés dans un grand lot, a donné les observations suivantes:

392 396 386 389 388 387 403 397 401 391 400 402 394 406 406 400

On considère que la résistance ohmique présente approximativement une distribution normale.

(a) Peut-on affirmer, au seuil de signification $\alpha = 0{,}05$, que le lot respecte la norme de 400 ohms?

(b) Calculer la probabilité de commettre une erreur de deuxième espèce avec le test effectué en (a) (c'est-à-dire avec les constantes de rejet pour $\bar{x}$ calculées en (a)) si la résistance ohmique présente en fait une distribution $\mathrm{N}(405, 49)$.

(c) Sous les mêmes hypothèses qu'en (b), de combien d'observations doit-on disposer pour que la probabilité de commettre une erreur de deuxième espèce soit inférieure ou égale à 0,05?

Solution. (a) On veut tester

$$H_0\colon \mu = 400 \quad \text{contre} \quad H_1\colon \mu \neq 400$$

au seuil $\alpha = 0{,}05$. On rejette H_0 si et seulement si

$$\bar{x} < 400 - t_{0{,}025;16-1}\frac{s}{\sqrt{16}} \quad \text{ou} \quad \bar{x} > 400 + t_{0{,}025;16-1}\frac{s}{\sqrt{16}}$$

On trouve que

$$\bar{x} = 396{,}125, \quad s \simeq 6{,}74 \quad \text{et} \quad t_{0{,}025;15} \overset{\text{tab. 5.2}}{\simeq} 2{,}131$$

Étant donné que

$$396{,}125 < 400 - 2{,}131\,\frac{6{,}74}{4} \simeq 396{,}41$$

on rejette l'hypothèse nulle H_0 au seuil de signification $\alpha = 0{,}05$.

(b) On cherche

$$P\left[396{,}41 \leq \bar{X} \leq 403{,}59 \mid \mu = 405\right], \quad \text{où } \bar{X} \sim \mathrm{N}\left(\mu, \tfrac{49}{16}\right)$$
$$\simeq P\left[-4{,}91 \leq \mathrm{N}(0,1) \leq -0{,}81\right] \simeq \Phi(-0{,}81) \overset{\text{tab. A.3}}{\simeq} 0{,}209$$

(c) On a:

$$n = \frac{(z_{0{,}025} + z_{0{,}05})^2\, 49}{(405 - 400)^2} \overset{\text{tab. 5.1}}{\simeq} (1{,}960 + 1{,}645)^2\,\frac{49}{25} \simeq 25{,}5$$

Donc, il faut disposer d'au moins 26 observations.

Question n$^{\text{o}}$ 18

On utilise deux machines identiques pour fabriquer une certaine pièce. On veut déterminer si les deux machines possèdent la même variabilité quant à une

caractéristique importante de cette pièce. Des échantillons aléatoires particuliers, prélevés de la production de chaque machine, ont donné les résultats suivants:

Machine A	140 135 140 138 135 138 140
Machine B	135 138 136 140 138 135 139

On a: $\bar{x}_A = 138$, $s_A \simeq 2{,}24$, $\bar{x}_B = 137{,}29$ et $s_B \simeq 1{,}98$.

(a) Préciser les hypothèses que l'on doit énoncer pour pouvoir effectuer le test.

(b) Effectuer le test statistique requis en utilisant un seuil de signification de 0,10. Que peut-on conclure?

(c) Tester l'hypothèse que les 14 données recueillies proviennent d'une distribution uniforme sur l'intervalle $[135, 140]$. Utiliser un seuil de signification $\alpha = 0{,}05$ et les intervalles suivants: $[135; 137{,}5)$ et $[137{,}5; 140]$.

Solution. (a) Soit X_A et X_B les deux variables aléatoires d'intérêt. On doit supposer que X_A et X_B présentent toutes les deux une distribution normale et qu'elles sont indépendantes.

(b) On veut tester

$$H_0\colon \sigma_A^2 = \sigma_B^2 \quad \text{contre} \quad H_1\colon \sigma_A^2 \neq \sigma_B^2$$

au seuil $\alpha = 0{,}10$. On rejette H_0 si et seulement si

$$f_0 > F_{0{,}05;7-1,7-1} \overset{\text{tab. } 5.4}{\simeq} 4{,}28 \quad \text{ou} \quad f_0 < F_{0{,}95;7-1,7-1} \simeq \frac{1}{4{,}28} \simeq 0{,}23$$

où

$$f_0 = \frac{s_A^2}{s_B^2} \simeq \left(\frac{2{,}24}{1{,}98}\right)^2 \simeq 1{,}28$$

Donc, on ne rejette pas H_0 au seuil de signification $\alpha = 0{,}10$; c'est-à-dire que l'on peut conclure que les deux machines ont la même variabilité.

(c) Soit X une observation quelconque. On veut tester

$$H_0\colon X \sim \mathrm{U}\,[135, 140] \quad \text{contre} \quad H_1\colon X \text{ n'est pas uniforme sur } [135, 140]$$

au seuil $\alpha = 0{,}05$. On a:

Intervalle	$[135; 137{,}5)$	$[137{,}5; 140]$	$\sum$
n_i	5	9	$14 = n$
p_i	$1/2$	$1/2$	1
np_i	7	7	14

La condition $m_i = np_i \geq 5$ est respectée pour $i = 1, 2$. On calcule alors

$$d^2 = \frac{(5-7)^2}{7} + \frac{(9-7)^2}{7} = \frac{8}{7} \simeq 1{,}14$$

On rejette H_0 si et seulement si

$$d^2 > \chi^2_{0{,}05;2-1} \overset{\text{tab. } 5.3}{\simeq} 3{,}84$$

Donc, on ne rejette pas H_0 au seuil de signification $\alpha = 0{,}05$.

Question n$^{\text{o}}$ 19

Soit X_1 et X_2 deux variables aléatoires indépendantes. On suppose que X_i présente une distribution de Poisson de paramètre λ_i, pour $i = 1, 2$.

(a) On peut montrer que $P[X_1 = k \mid X_1 + X_2 = n] = \binom{n}{k}\theta^k (1-\theta)^{n-k}$ pour $k = 0, 1, \ldots, n$, où $\theta := \frac{\lambda_1}{\lambda_1+\lambda_2}$. Pour tester

$$H_0\colon \lambda_1 = \lambda_2 \quad \text{contre} \quad H_1\colon \lambda_1 > \lambda_2$$

on prélève une observation X_{i1} de chaque variable aléatoire et on rejette H_0 si et seulement si X_{11}, étant donné que $X_{11} + X_{21} = n$, est supérieure à une constante C. Utiliser l'approximation de la distribution binomiale par la distribution normale (sans correction de continuité) pour montrer que la constante C qu'il faut choisir, pour obtenir un test avec un risque de première espèce (c'est-à-dire un seuil de signification) de α, est donnée par

$$C \simeq \frac{n}{2} + z_\alpha \frac{\sqrt{n}}{2} \quad \text{(si } n \text{ est assez grand)}$$

(b) Calculer la valeur de β pour le test en (a) si $n = 100$, $\alpha = 0{,}05$, $\lambda_1 = 60$ et $\lambda_2 = 40$.

(c) Calculer la constante C en (a) si l'on dispose en fait d'un échantillon aléatoire $X_{11}, \ldots, X_{1k}$ de X_1 et d'un échantillon aléatoire $X_{21}, \ldots, X_{2m}$ de X_2, et si $\sum_{i=1}^{k} X_{1i} + \sum_{j=1}^{m} X_{2j} = n$.

Solution. (a) On a:

$$P[X_{11} > C \mid X_{11} + X_{12} = n] = P[\mathrm{B}(n, \theta) > C]$$

Alors, si n est assez grand, on peut écrire, sous H_0, que la constante C doit être telle que

$$P\left[\mathrm{N}\left(\tfrac{n}{2},\tfrac{n}{4}\right)>C\right]\simeq\alpha\iff\frac{C-\frac{n}{2}}{\sqrt{n}/2}\simeq z_\alpha\iff C\simeq\frac{n}{2}+z_\alpha\frac{\sqrt{n}}{2}$$

(b) On calcule

$$C\simeq\frac{100}{2}+z_{0,05}\frac{\sqrt{100}}{2}\simeq 58{,}225$$

Alors on peut écrire que

$$\begin{aligned}\beta\left(\theta=\frac{60}{60+40}\right)&\simeq P\left[\mathrm{N}\left(n\theta,n\theta\left(1-\theta\right)\right)\leq C\right]\\&\simeq P\left[\mathrm{N}(0,1)\leq\frac{58{,}225-60}{\sqrt{24}}\right]\simeq\Phi(-0{,}36)\overset{\text{tab. A.3}}{\simeq}0{,}36\end{aligned}$$

(c) On a: $\sum_{i=1}^{k}X_{1i}\sim\mathrm{Poi}(\lambda_{X_1}=k\lambda_1)$ et $\sum_{j=1}^{m}X_{2j}\sim\mathrm{Poi}(\lambda_{X_2}=m\lambda_2)$. Alors on peut écrire que

$$\sum_{i=1}^{k}X_{1i}\;\Bigg|\;\left\{\sum_{i=1}^{k}X_{1i}+\sum_{j=1}^{m}X_{2}j=n\right\}\sim\mathrm{B}\left(n,\theta:=\frac{k\lambda_1}{k\lambda_1+m\lambda_2}\right)$$

Il s'ensuit que

$$P\left[\sum_{i=1}^{k}X_{1i}>C\;\Bigg|\;\sum_{i=1}^{k}X_{1i}+\sum_{j=1}^{m}X_{2j}=n\right]\overset{H_0}{=}P\left[\mathrm{B}\left(n,\theta_0\right)>C\right]$$

où $\theta_0:=k/(k+m)$. Si n est grand, alors on peut écrire que

$$P\left[\mathrm{N}(n\theta_0,n\theta_0(1-\theta_0))>C\right]\simeq\alpha\iff C\simeq n\theta_0+z_\alpha\sqrt{n\theta_0(1-\theta_0)}$$

Question n° 20

Pour vérifier si un dé est bien équilibré, on le lance 48 fois. On observe les résultats suivants:

Face	1	2	3	4	5	6	Total
Effectif	5	11	7	6	7	12	48

(a) Tester l'hypothèse que le dé est bien équilibré au seuil de signification $\alpha=0{,}05$.

(b) Avec ces mêmes données, on veut vérifier si la face "6" a plus d'une chance sur six d'apparaître. Quelle est la conclusion du test si l'on prend $\alpha=0{,}10$?

(c) Pour le test utilisé en (b), donner la taille de l'échantillon aléatoire nécessaire pour obtenir une valeur de β de 5 % si, en réalité, la face "6" a 20 % des chances d'apparaître.

(d) Supposons que l'on dispose plutôt du tableau d'effectifs suivant:

Face	1	2	3	4	5	6	Total
Effectif	k_1	k_2	7	6	7	12	48

Pour quelles valeurs de $k_1^2 + k_2^2$ rejette-t-on l'hypothèse selon laquelle le dé est bien équilibré (si l'on prend $\alpha = 0{,}05$)?

Solution. (a) On veut tester

$$H_0\text{: } p_i = 1/6 \ \ \forall i \quad \text{contre} \quad H_1\text{: } p_i \text{ n'est pas identique à } 1/6$$

où $p_i := P\,[\text{Face} = i]$ pour $i = 1, \ldots, 6$, au seuil de signification $\alpha = 0{,}05$. On a:

i	1	2	3	4	5	6	$\sum$
n_i	5	11	7	6	7	12	$n = 48$
p_i	1/6	1/6	1/6	1/6	1/6	1/6	1
$np_i = m_i$	8	8	8	8	8	8	48

Notons que $m_i \equiv 8 > 5$. On calcule alors

$$d^2 = \frac{(5-8)^2}{8} + \ldots + \frac{(12-8)^2}{8} = \frac{40}{8} = 5$$

On rejette H_0 si et seulement si

$$d^2 > \chi^2_{0{,}05;6-0-1} \overset{\text{tab. 5.3}}{\simeq} 11{,}07$$

Donc, on ne rejette pas H_0 au seuil $\alpha = 0{,}05$.

Remarque. On a utilisé un test d'ajustement du khi-deux de Pearson plutôt que le test de l'égalité de plusieurs proportions ici, car on désire tester l'hypothèse

$$H_0\text{: } p_1 = p_2 = \ldots = p_6 = 1/6$$

et non pas

$$H_0\text{: } p_1 = p_2 = \ldots = p_6$$

Ainsi, il n'est pas nécessaire d'estimer la proportion commune p, comme dans le cas du test de l'égalité de plusieurs proportions.

(b) On veut maintenant tester

$$H_0\colon p = 1/6 \quad \text{contre} \quad H_1\colon p > 1/6$$

où $p := P\,[\text{Face} = 6]$, au seuil de signification $\alpha = 0{,}10$. On calcule

$$\hat{p} = \frac{12}{48} = 0{,}25$$

On rejette H_0 si et seulement si

$$\hat{p} > \frac{1}{6} + \underbrace{z_{0{,}10}}_{1{,}282} \sqrt{\frac{(1/6)\,(5/6)}{48}} \simeq 0{,}2356$$

Donc, on rejette H_0 au seuil $\alpha = 0{,}10$.

(c) On doit prendre

$$n \simeq \left(\frac{z_{0{,}10}\sqrt{(1/6)\,(5/6)} + z_{0{,}05}\sqrt{(0{,}2)\,(0{,}8)}}{0{,}2 - \frac{1}{6}} \right)^2 \overset{\text{tab. 5.1}}{\simeq} 1161$$

(d) On calcule maintenant

$$d^2 = \frac{(k_1 - 8)^2}{8} + \frac{(k_2 - 8)^2}{8} + \ldots + \frac{(12 - 8)^2}{8} = \frac{(k_1 - 8)^2 + (k_2 - 8)^2 + 22}{8}$$

Puisque $k_1 + k_2 = 48 - 32 = 16$, on peut écrire que

$$d^2 = \frac{1}{8}\left[k_1^2 + k_2^2 - 16(k_1 + k_2) + 150\right] = \frac{1}{8}\left(k_1^2 + k_2^2 - 106\right)$$

Étant donné que l'on rejette H_0 au seuil de signification $\alpha = 0{,}05$ si et seulement si $d^2 > 11{,}07$, on conclut que l'on rejette l'hypothèse selon laquelle le dé est bien équilibré si et seulement si

$$k_1^2 + k_2^2 > 194{,}56 \quad \text{(environ)}$$

Question n$^{\text{o}}$ 21

On s'intéresse à la résistance des tiges métalliques de deux entreprises. On dispose de deux lots de tiges provenant de chacune des entreprises. On tire de

façon aléatoire 10 tiges de chaque lot et on mesure la tension nécessaire pour briser chacune d'elles. On obtient les résultats suivants:

Lot 1	53,2 53,9 53,1 50,9 42,2 52,9 55,8 41,9 50,0 49,0
Lot 2	55,0 64,5 53,0 57,8 56,1 58,0 55,8 50,8 54,0 52,2

On suppose que la résistance des tiges métalliques présente approximativement une distribution normale.

(a) Tester l'hypothèse que la résistance moyenne est la même dans les deux lots, au seuil de signification $\alpha = 0{,}05$.

Indication. On a: $t_{0{,}025;18} \simeq 2{,}101$.

(b) Tester, en supposant que les 20 tiges proviennent en fait de la même entreprise, l'hypothèse que (i) la résistance moyenne d'une tige métallique est de 55 et (ii) la variance de la résistance d'une tige métallique est de 20. Utiliser $\alpha = 0{,}05$.

Indication. On a: $t_{0{,}025;19} \simeq 2{,}093$, $\chi^2_{0{,}025;19} \simeq 32{,}85$ et $\chi^2_{0{,}975;19} \simeq 8{,}91$.

Solution. (a) Soit X_i la résistance d'une tige métallique de l'entreprise i, pour $i = 1, 2$. On suppose que $X_i \sim \mathrm{N}\big(\mu_i, \sigma_i^2\big)$ et que X_1 et X_2 sont indépendantes. On teste d'abord

$$H_0\colon \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad H_1\colon \sigma_1^2 \neq \sigma_2^2$$

au seuil $\alpha = 0{,}05$. On trouve que

$$s_1 \simeq 4{,}767 \quad \text{et} \quad s_2 \simeq 3{,}862$$

On calcule

$$f_0 = \left(\frac{s_1}{s_2}\right)^2 \simeq 1{,}52$$

On rejette H_0 si et seulement si

$$f_0 > F_{0{,}025;9,9} \overset{\text{tab. } 5.4}{\simeq} 4{,}03 \quad \text{ou} \quad f_0 < F_{0{,}975;9,9} \simeq \frac{1}{4{,}03} \simeq 0{,}25$$

Donc, on ne rejette pas l'égalité des variances au seuil $\alpha = 0{,}05$. On suppose alors que $\sigma_1 = \sigma_2 := \sigma$. On veut tester

$$H_0\colon \mu_1 = \mu_2 \quad \text{contre} \quad H_1\colon \mu_1 \neq \mu_2$$

au seuil $\alpha = 0{,}05$. On trouve que

$$\bar{x}_1 \simeq 50{,}29 \quad \text{et} \quad \bar{x}_2 \simeq 55{,}72$$

On calcule

$$s_p^2 = \frac{(10-1)\,s_1^2 + (10-1)\,s_2^2}{10+10-2} = \frac{s_1^2 + s_2^2}{2} \simeq 18{,}82$$

On rejette H_0 si et seulement si

$$|\bar{x}_1 - \bar{x}_2| > \underbrace{t_{0{,}025;10+10-2}}_{2{,}101}\, s_p \sqrt{\frac{1}{10} + \frac{1}{10}} \simeq 4{,}08$$

Puisque $\bar{x}_1 - \bar{x}_2 \simeq -5{,}43$, on rejette H_0 au seuil de signification $\alpha = 0{,}05$.

(b) Soit X la résistance d'une tige métallique. On suppose que $X \approx \mathrm{N}(\mu, \sigma^2)$.

(i) On veut tester, au seuil $\alpha = 0{,}05$,

$$H_0\colon \mu = 55 \quad \text{contre} \quad H_1\colon \mu \neq 55$$

On trouve que $\bar{x} \simeq 53{,}0$ et $s \simeq 5{,}1$. On rejette H_0 si et seulement si

$$\bar{x} < C_1 = 55 - \underbrace{t_{0{,}025;19}}_{2{,}093} \frac{s}{\sqrt{20}} \simeq 52{,}6$$

ou

$$\bar{x} > C_2 \simeq 55 + 2{,}093 \frac{s}{\sqrt{20}} \simeq 57{,}4$$

Donc, on ne rejette pas H_0 si $\alpha = 0{,}05$.

(ii) On veut finalement tester

$$H_0\colon \sigma^2 = 20 \quad \text{contre} \quad H_1\colon \sigma^2 \neq 20$$

au seuil $\alpha = 0{,}05$, en utilisant les 20 données. On trouve que $s^2 \simeq 25{,}59$. On rejette H_0 si et seulement si

$$\frac{19}{20} s^2 < \chi^2_{0{,}975;19} \simeq 8{,}91 \quad \text{ou} \quad \frac{19}{20} s^2 > \chi^2_{0{,}025;19} \simeq 32{,}85$$

Puisque $\frac{19}{20} s^2 \simeq 24{,}31$, dans ce cas non plus on ne rejette pas H_0 si $\alpha = 0{,}05$.

Question n$^{\text{o}}$ 22

On étudie l'efficacité thermique X (en %) des moteurs Diesel fabriqués par un grand constructeur d'automobiles. On suppose que la variable aléatoire X

présente (approximativement) une distribution normale dont l'écart-type est égal à 2. Des tests effectués sur 25 moteurs ont donné les résultats suivants: $\bar{x} = 31{,}4$ et $s = 1{,}6$.

(a) Tester l'hypothèse H_0: $\mu = 32{,}3$ contre H_1: $\mu \neq 32{,}3$. Utiliser $\alpha = 0{,}05$.

(b) Calculer la probabilité de commettre une erreur de deuxième espèce en (a) si l'efficacité thermique moyenne est en fait de 31,3 %.

(c) En se basant sur les données recueillies, peut-on conclure que l'écart-type de X est en fait inférieur à 2 %? Effectuer un test d'hypothèses au seuil de signification $\alpha = 0{,}01$.

Indication. On a: $\chi^2_{0{,}99;24} \simeq 10{,}86$.

Solution. (a) On veut tester

$$H_0\colon \mu = 32{,}3 \quad \text{contre} \quad H_1\colon \mu \neq 32{,}3$$

au seuil de signification $\alpha = 0{,}05$. On rejette H_0 si et seulement si

$$\bar{x} < C_1 = 32{,}3 - z_{0{,}025}\frac{2}{\sqrt{25}} \simeq 31{,}516$$

ou

$$\bar{x} > C_2 \simeq 32{,}3 + 1{,}960\frac{2}{\sqrt{25}} \simeq 33{,}084$$

Puisque $\bar{x} = 31{,}4 < 31{,}516$, on rejette H_0 au seuil $\alpha = 0{,}05$.

(b) On cherche

$$\begin{aligned} \beta\,(\mu = 31{,}3) &\simeq P\left[31{,}516 \leq \bar{X} \leq 33{,}084 \mid \mu = 31{,}3\right] \\ &\simeq P\left[31{,}516 \leq \mathrm{N}\big(31{,}3; \tfrac{4}{25}\big) \leq 33{,}084\right] = P\left[0{,}54 \leq \mathrm{N}(0,1) \leq 4{,}46\right] \\ &\overset{\text{tab. A.3}}{\simeq} 1 - 0{,}7054 = 0{,}2946 \end{aligned}$$

(c) On veut tester

$$H_0\colon \sigma^2 = 4 \quad \text{contre} \quad H_1\colon \sigma^2 < 4$$

au seuil $\alpha = 0{,}01$. On rejette H_0 si et seulement si

$$\frac{(25-1)}{4}s^2 < \chi^2_{1-0{,}01;25-1} \simeq 10{,}86$$

Puisque

$$s^2 = (1{,}6)^2 = 2{,}56 \implies 6s^2 = 15{,}36 \geq 10{,}86$$

on ne peut pas rejeter H_0 au seuil $\alpha = 0{,}01$.

Question n^o 23

Des mesures de pourcentage d'élongation ont été effectuées sur 10 pièces d'acier. Cinq de ces pièces furent traitées en utilisant la méthode A (aluminium seulement) et les cinq autres en utilisant la méthode B (aluminium plus calcium), avec les résultats suivants:

Méthode A (%)	28 29 31 33 30
Méthode B (%)	34 27 30 36 33

(a) Tester l'égalité des variances. Utiliser $\alpha = 0{,}10$.

(b) Peut-on conclure que les deux méthodes donnent, en moyenne, les mêmes résultats? Utiliser $\alpha = 0{,}05$.

Solution. (a) Soit X_A (respectivement X_B) le pourcentage d'élongation des pièces traitées en utilisant la méthode A (respectivement B). On suppose que $X_A \sim \mathrm{N}(\mu_A, \sigma_A^2)$ et $X_B \sim \mathrm{N}(\mu_B, \sigma_B^2)$ sont indépendantes. On veut tester

$$H_0\colon \sigma_A^2 = \sigma_B^2 \quad \text{contre} \quad H_1\colon \sigma_A^2 \neq \sigma_B^2$$

au seuil $\alpha = 0{,}10$. On rejette H_0 si et seulement si

$$f_0 = \frac{s_A^2}{s_B^2} < F_{1-0{,}05;5-1,5-1} \quad \text{ou} \quad f_0 > F_{0{,}05;4,4} \overset{\text{tab. 5.4}}{\simeq} 6{,}39$$

Puisque

$$s_A^2 \simeq (1{,}92)^2 \simeq 3{,}70, \quad s_B^2 \simeq (3{,}54)^2 \simeq 12{,}50$$

et

$$F_{0{,}95;4,4} \simeq \frac{1}{6{,}39} \simeq 0{,}16$$

on a: $f_0 \simeq 0{,}30$ et

$$0{,}16 \leq f_0 \leq 6{,}39$$

Donc, on ne rejette pas H_0 au seuil de signification $\alpha = 0{,}10$.

(b) On veut maintenant tester

$$H_0\colon \mu_A = \mu_B \quad \text{contre} \quad H_1\colon \mu_A \neq \mu_B$$

au seuil $\alpha = 0{,}05$. On suppose que $\sigma_A = \sigma_B$ (car la conclusion en (a) est la même au seuil $\alpha = 0{,}05$) et on pose:

$$s_p^2 = \frac{(5-1)\,s_A^2 + (5-1)\,s_B^2}{5+5-2} = \frac{s_A^2 + s_B^2}{2} \simeq 8{,}1$$

On rejette H_0 si et seulement si

$$\left| \frac{30{,}2 - 32{,}0}{\sqrt{8{,}1}\sqrt{\frac{1}{5} + \frac{1}{5}}} \right| \simeq 1 > t_{0{,}025;5+5-2} \overset{\text{tab. } 5.2}{\simeq} 2{,}306$$

Donc, on ne rejette pas H_0 si $\alpha = 0{,}05$.

Question n$^{\text{o}}$ 24

Une machine fabrique des billes en métal que l'on classe (en fonction de leur diamètre) en trois catégories: A, B et C. Une bille est dite de catégorie A si son diamètre est inférieur à 9,5 mm, de catégorie B si son diamètre est au moins de 9,5 mm et au plus de 10,5 mm, et de catégorie C si son diamètre est supérieur à 10,5 mm.

Des études antérieures ont montré que 80 % des billes provenant de la machine étaient de catégorie B et qu'il y avait trois fois plus de billes de catégorie C que de catégorie A. Un échantillon aléatoire particulier de 200 billes est prélevé de la production de la machine et on y observe 12 billes de catégorie A, 156 de catégorie B et 32 de catégorie C.

Effectuer un test d'ajustement du khi-deux de Pearson pour vérifier l'hypothèse que la répartition des billes selon les trois catégories n'a pas changé depuis la dernière étude. Utiliser $\alpha = 0{,}01$.

Solution. Soit X la variable aléatoire qui prend la valeur 1 si la bille est de catégorie A, la valeur 2 si la bille est de catégorie B, et la valeur 3 si elle est de catégorie C. On veut tester

$$H_0\colon p_X(x) = p_0(x) \quad \text{contre} \quad H_1\colon p_X(x) \neq p_0(x)$$

au seuil $\alpha = 0{,}01$, où

$$p_0(1) = 0{,}05, \quad p_0(2) = 0{,}80 \quad \text{et} \quad p_0(3) = 0{,}15$$

(de sorte que $p_0(3) = 3\,p_0(1)$ et $p_0(1) + p_0(2) + p_0(3) = 1$). On a:

i	1	2	3	$\sum$
n_i	12	156	32	$200 = n$
p_i	0,05	0,80	0,15	1
$np_i = m_i$	10	160	30	200

La condition $m_i = np_i \geq 5\ \forall i$ est respectée. On calcule ensuite

$$d^2 = \frac{(12-10)^2}{10} + \frac{(156-160)^2}{160} + \frac{(32-30)^2}{30} \simeq 0{,}633$$

Puisque

$$d^2 \leq \chi^2_{0{,}01;3-0-1} \overset{\text{tab. } 5.3}{\simeq} 9{,}21$$

le modèle est accepté au seuil $\alpha = 0{,}01$.

Exercices

Question n° 1

Dans une certaine réaction chimique, il est très important qu'une solution qu'on désire utiliser comme réactif ait un pH de 8,30. On sait que la méthode qu'on utilise pour déterminer le pH des solutions de ce type donne des mesures qui présentent une distribution approximativement normale, dont la moyenne est la vraie valeur du pH et l'écart-type σ est égal à 0,02. Si le pH de la solution est vraiment égal à 8,30, on veut que le test que l'on utilisera arrive à cette conclusion avec une probabilité de 0,95. De plus, on veut que la probabilité de détecter une différence de 0,03 entre la vraie valeur du pH et le 8,30 souhaité soit d'au moins 0,95.

(a) Quel est le nombre n d'observations que l'on doit recueillir?

(b) Si la moyenne d'un échantillon aléatoire particulier de taille n (calculé en (a)) est égale à 8,31, quelle est la conclusion du test?

(c) Quelle est, en fait, la probabilité de détecter une différence de 0,03 avec le test qu'on a construit?

Question n° 2

On s'intéresse au temps pendant lequel un guichet automatique est utilisé au cours d'une journée, entre 7 h et 23 h. Un étude du temps d'utilisation du guichet au cours des 100 derniers jours a permis de constituer le tableau suivant:

Nombre d'heures	$[0,4)$	$[4,8)$	$[8,12)$	$[12,16]$
Nombre de jours	20	30	30	20

(a) Tester, au seuil de signification $\alpha = 0{,}05$, l'ajustement des distributions suivantes aux données:

(i) uniforme sur l'intervalle $[0, 16]$;

(ii) exponentielle de paramètre $\lambda = 1/8$;

(iii) normale de moyenne $\mu = 8$ et variance $\sigma^2 = 16$.

(b) Parmi les trois distributions en (a), laquelle est la seule dont peuvent réellement provenir les données? Justifier votre réponse.

Question n^{o} 3

Un fabricant d'automobiles prétend que la consommation moyenne de ses véhicules ne dépasse pas 6,0 litres par 100 kilomètres. Soit X la variable aléatoire qui représente la consommation (en litres aux 100 kilomètres) des véhicules en question. On suppose que X présente approximativement une distribution normale d'écart-type 0,3. On a obtenu l'échantillon aléatoire particulier suivant de X:

$$6,3 \quad 5,7 \quad 5,9 \quad 6,1 \quad 6,0 \quad 5,8 \quad 6,5 \quad 6,6 \quad 5,9 \quad 6,2$$

(a) Peut-on rejeter la prétention du fabricant au seuil de signification $\alpha = 0{,}01$ en se basant sur les données recueillies?

(b) Calculer le risque de deuxième espèce du test utilisé en (a) si la moyenne de X est en fait égale à 6,3 l/100 km.

(c) Calculer le nombre d'observations que l'on doit recueillir, si l'on veut que la probabilité de commettre une erreur de deuxième espèce ne soit pas plus grande que 0,10 lorsque $E[X] = 6{,}3$ et $\alpha = 0{,}01$.

(d) Obtenir un intervalle de confiance avec un coefficient de confiance de 99 % pour la moyenne de X.

Question n^{o} 4

On étudie la durée de vie (en heures) d'un nouveau type d'ampoule. Un échantillon aléatoire particulier de 50 ampoules a permis de constituer le tableau d'effectifs suivant:

Durée	$[0, 100)$	$[100, 200)$	$[200, 300)$	$[300, 400)$	$[400, 500)$	$[500, \infty)$
Nombre	3	10	15	12	6	4

De plus, la moyenne des 50 données est de 275.

(a) Effectuer un test d'ajustement du khi-deux de Pearson d'une distribution normale, de moyenne inconnue et d'écart-type $\sigma = 150$, aux données. Utiliser un seuil de signification de 0,10.

(b) Soit θ la proportion des ampoules qui durent au moins 400 heures. Tester

$$H_0\colon \theta = 0{,}25 \quad \text{contre} \quad H_1\colon \theta < 0{,}25$$

au seuil de signification $\alpha = 0{,}10$.

Question n$^{\text{o}}$ 5

On a chronométré le temps de service (en minutes) à un comptoir d'information et produit le tableau d'effectifs suivant:

Temps	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$	$[5,\infty)$	Total
Nbre de clients	53	36	28	12	10	11	150

De plus, on a: $\sum_{i=1}^{150} t_i = 300$, où t_i est le temps de service observé pour le i^{e} client. Tester, au seuil de signification $\alpha = 0{,}05$, l'hypothèse que le temps de service est une variable aléatoire qui présente une distribution exponentielle.

Question n$^{\text{o}}$ 6

Une étude a été effectuée pour comparer le rendement (mesuré en %) d'une réaction chimique en utilisant deux catalyseurs, A et B. Les données sont:

$$\text{Catalyseur } A: \ 86,\ 83,\ 85,\ 84,\ 82,\ 84$$
$$\text{Catalyseur } B: \ 85,\ 89,\ 87,\ 88,\ 89,\ 88,\ 90$$

Formuler le problème statistique sous la forme de tests d'hypothèses, préciser la notation utilisée et toutes les *hypothèses de base*. Utiliser un seuil de signification de 0,05 pour effectuer les tests d'hypothèses.

Question n$^{\text{o}}$ 7

Il y a environ 40.000 collisions arrière par année pour les 10 millions d'automobiles au Canada. On se demande si un troisième feu d'indication d'arrêt, situé plus haut que les deux habituels et ne créant aucune confusion avec les feux de position de nuit, réduirait la fréquence de tels accidents. On décide d'installer ce troisième feu sur un certain nombre de véhicules, à titre d'expérience, pendant un an.

Soit θ la probabilité de collision arrière; actuellement, $\theta = \theta_0 = 0{,}004$ (40.000/10.000.000). En considérant les coûts supplémentaires d'un tel feu et les coûts sociaux de collisions, on juge que la collectivité serait gagnante si de tels accidents diminuaient de 30 % avec le troisième feu d'arrêt. On voudrait donc rejeter avec une probabilité de 0,99 l'hypothèse nulle $H_0\colon \theta = \theta_0$ lorsque $\theta = 0{,}7\,\theta_0$. Le seuil de signification du test est fixé à 0,05.

(a) Déterminer le nombre de voitures qui devront participer à l'expérience. On utilisera une approximation de la distribution binomiale par une distribution normale.

(b) Déterminer la région critique (région de rejet de H_0) en nombre d'accidents au cours d'une année pour les véhicules participant à l'expérience.

(c) Étant donné que l'expérience décrite en (a) et (b) est impraticable (trop grand nombre de voitures), il est décidé de considérer un échantillon de 2000 automobiles. Dans ce cas, l'approximation normale du nombre X d'accidents étant moins bonne, on approche plutôt la distribution de X par une distribution de Poisson de paramètre $\lambda = n\theta_0 = 8$. On propose le test suivant: on rejette H_0 si et seulement si $X \leq C$. Déterminer la valeur critique C pour un seuil de signification $\alpha = 0{,}05$.

(d) Calculer le risque de deuxième espèce du test effectué en (c) lorsque $\theta = 0{,}5\,\theta_0$.

Question n$^{\text{o}}$ 8

Les données qui suivent représentent le résultat d'une enquête menée auprès des étudiant(e)s de quatre universités et visant à déterminer la proportion de fumeurs parmi les étudiant(e)s:

Statut \ Université	1	2	3	4
Fumeurs	10	20	20	10
Non-fumeurs	30	30	40	40

(a) Y a-t-il une relation entre les variables Université et Statut? Utiliser $\alpha = 0{,}05$.

(b) Obtenir un intervalle de confiance à 95 % pour le pourcentage de fumeurs parmi les étudiant(e)s de l'université n$^{\text{o}}$ 4.

(c) On suppose que les quatre universités ont approximativement le même nombre d'étudiant(e)s. Tester, au seuil de signification $\alpha = 0{,}05$, l'hypothèse nulle H_0: les étudiant(e)s interrogé(e)s ont été pris(es) au hasard parmi l'ensemble des étudiant(e)s des quatre universités.

Question n$^{\text{o}}$ 9

Soit X_1 et X_2 les variables aléatoires désignant le diamètre (en millimètres) de fils métalliques avant et après une modification du procédé de fabrication. On dispose des deux échantillons aléatoires particuliers suivants:

$$\begin{array}{ll} X_1\text{:} & 1{,}4\ \ 1{,}8\ \ 1{,}5\ \ 1{,}7\ \ 2{,}0 \\ X_2\text{:} & 1{,}5\ \ 1{,}7\ \ 2{,}1\ \ 1{,}8\ \ 1{,}6\ \ 1{,}8 \end{array}$$

(a) Quelles hypothèses doit-on faire au sujet des variables aléatoires X_1 et X_2 pour pouvoir effectuer des tests d'hypothèses concernant les moyennes et les variances de ces deux variables?

(b) Peut-on affirmer que la modification du procédé de fabrication a réduit la variance? Utiliser $\alpha = 0{,}01$.

(c) Tester l'égalité des moyennes au seuil de signification $\alpha = 0{,}05$.

Question n^{o} 10

Un tour automatique produit des axes dont on arrive à contrôler le diamètre moyen. Le procédé de fabrication est sous contrôle statistique lorsque l'écart-type du diamètre ne dépasse pas 1,2 mm. En cours de production, on veut tester l'hypothèse nulle que le procédé est sous contrôle, avec un risque de première espèce de 0,05.

(a) Écrire l'hypothèse nulle et la contre-hypothèse, et préciser les hypothèses de base utilisées.

(b) Calculer la région critique, en fonction de S^2, pour un échantillon de taille 10.

(c) Calculer le risque de deuxième espèce du test en (b), si l'écart-type véritable est de 2,1 mm.

(d) Donner la conclusion du test si l'on a obtenu les 10 observations particulières suivantes:

23,8	25,7	26,5	22,3	24,9
27,6	24,7	23,2	26,8	24,1

(e) Calculer la taille de l'échantillon requise pour que le risque de deuxième espèce soit de 0,10, si l'écart-type est de 2,1 mm.

Question n^{o} 11

La finale entre les équipes A et B, dans la Ligue nationale de hockey (LNH), se joue en au plus sept matchs (sans match nul). L'équipe qui gagne quatre matchs est déclarée gagnante. On énonce les hypothèses suivantes:

- les matchs sont indépendants les uns des autres;
- la probabilité que l'équipe A gagne un match est de θ;
- l'équipe A est au moins aussi bonne que l'équipe B: $0{,}5 \leq \theta < 1$.

Si X est le nombre de matchs requis pour déterminer le gagnant, alors la fonction de probabilité de X est:

$$p_X(k;\theta) = \begin{cases} \theta^4 + (1-\theta)^4 & \text{si } k = 4 \\ 4\theta(1-\theta)\left[\theta^3 + (1-\theta)^3\right] & \text{si } k = 5 \\ 10\theta^2(1-\theta)^2\left[\theta^2 + (1-\theta)^2\right] & \text{si } k = 6 \\ 20\theta^3(1-\theta)^3 & \text{si } k = 7 \end{cases}$$

(a) Justifier la formule donnant $p_X\,(4;\theta)$.

(b) Quatre-vingt-trois finales de la LNH ont donné les résultats suivants:

k	4	5	6	7	Total
n_k	15	26	24	18	83

Tester, à l'aide d'un test du khi-deux de Pearson, l'ajustement du modèle proposé avec $\theta = 0{,}5$. Utiliser un seuil de signification de 0,05.

(c) On décide d'ajuster le modèle proposé en estimant θ par la méthode de vraisemblance maximale avec les données de la LNH. Comme la fonction de vraisemblance $L\,(\theta)$ est incommode à manipuler et comme l'équation pour déterminer l'estimateur à vraisemblance maximale, θ_{VM}, est difficile à résoudre, on a évalué numériquement $\ln L\,(\theta)$ avec $\theta = 0{,}50\ (0{,}05)\ 0{,}95$ ainsi que $p_X\,(k;\theta)$. Les calculs donnent:

θ	$\ln L\,(\theta)$	$p_X\,(4;\theta)$	$p_X\,(5;\theta)$	$p_X\,(6;\theta)$	$p_X\,(7;\theta)$
0,50	−116	0,1250	0,2500	0,3125	0,3125
0,55	−115	0,1325	0,2549	0,3093	0,3033
0,60	−114	0,1552	0,2688	0,2995	0,2765
0,65	−113	0,1935	0,2889	0,2821	0,2355
0,70	−114	0,2482	0,3108	0,2558	0,1852
0,75	−118	0,3203	0,3281	0,2197	0,1319
0,80	−128	0,4112	0,3328	0,1741	0,0819
0,85	−147	0,5225	0,3149	0,1211	0,0414
0,90	−182	0,6562	0,2628	0,0664	0,0146
0,95	−254	0,8145	0,1629	0,0204	0,0022

(i) Quel est, selon ce tableau, l'estimateur à vraisemblance maximale θ_{VM}?

(ii) Tester l'ajustement du modèle avec θ_{VM} (au seuil de signification $\alpha = 0{,}05$).

Question n$^{\text{o}}$ 12

On a effectué une étude pour décider si l'on doit adopter un nouveau procédé de liquéfaction du charbon. On croit que le nouveau procédé a une efficacité moyenne plus grande que l'ancien et que les variations sont plus petites. L'efficacité est mesurée par la quantité de combustible synthétique obtenue par kilogramme d'hydrogène utilisé. Les données de l'étude sont:

Ancien procédé				Nouveau procédé			
11,1	12,6	12,1	14,2	16,4	16,8	16,7	17,0
10,5	15,3	14,5	13,2	17,2	16,9	16,3	16,2
10,9	14,2	15,6	12,6	17,2	17,1		

(a) Comparer les variations de l'efficacité des deux procédés au moyen d'un test statistique approprié. Préciser la notation, les hypothèses de base employées, l'hypothèse nulle et la contre-hypothèse. Effectuer le test avec un seuil de signification de 0,05.

(b) Comparer l'efficacité moyenne des deux procédés au moyen d'un test qui tient compte du résultat obtenu en (a). Utiliser un seuil de signification $\alpha = 0{,}05$.

(c) Refaire la partie (a) si, dans le cas de l'ancien procédé, on dispose plutôt des résultats suivants: $n = 16$, $\bar{x} = 13{,}08$ et $s = 0{,}54$.

Indication. On a: $F_{0,05;15,9} \simeq 3{,}01$.

(d) Comparer l'efficacité moyenne des deux procédés au moyen d'un *intervalle de confiance* qui tient compte du résultat obtenu en (c). Utiliser un coefficient de confiance de 0,95.

Question n$^{\text{o}}$ 13

Un procédé de fabrication est sous contrôle statistique si les mesures qu'il génère proviennent d'une distribution $\mathrm{N}\big(\mu = 100, \sigma^2 = 100\big)$. Pour contrôler le procédé, on dispose de cinq règles de décision: D_1, D_2, D_3, D_4 et D_5, dont on se propose d'analyser l'efficacité. Périodiquement, par exemple toutes les heures, un échantillon aléatoire de taille 5 est obtenu et on calcule

$$\bar{X} = \tfrac{1}{5}\sum_{i=1}^{5} X_i,\ S^2 = \tfrac{1}{4}\sum_{i=1}^{5}\left(X_i - \bar{X}\right)^2,\ S^{*2} = \tfrac{1}{5}\sum_{i=1}^{5}\left(X_i - 100\right)^2$$

Les règles de décision sont définies dans le tableau qui suit:

	Le procédé est déclaré	
Règle	sous contrôle si	hors de contrôle
D_1	$95{,}31 < \bar{X} < 104{,}69$	autrement
D_2	$26{,}60 < S^2 < 194{,}47$	autrement
D_3	$32{,}20 < S^{*2} < 121{,}28$	autrement
D_4	D_1 et D_2 respectées	autrement
D_5	D_1 et D_3 respectées	autrement

(a) Pour chacune des règles (D_1 à D_5), calculer le risque (c'est-à-dire la probabilité) de déclarer le procédé hors de contrôle alors qu'il est sous contrôle.

Notons que l'on peut démontrer l'indépendance des variables aléatoires $\bar{X}$ et S^2 ainsi que $\bar{X}$ et S^{*2}.

Indication. On a: $\chi^2_{0,30;5} \simeq 6{,}06$.

(b) En se basant sur les calculs faits en (a), quelle règle de décision devrait-on choisir? Justifier votre réponse.

Question n$^{\text{o}}$ 14

La qualité de reproduction d'un photocopieur est mesurée en dénombrant les imperfections sur une page photocopiée. On admet que le nombre X d'imperfections présente approximativement une distribution de Poisson de paramètre λ. On veut tester l'hypothèse nulle H_0: $\lambda = \lambda_0$ contre H_1: $\lambda > \lambda_0$ à l'aide d'un échantillon aléatoire de taille n: $X_1, X_2, \ldots, X_n$.

(a) Justifier le résultat suivant:

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}}$$

présente approximativement une distribution $\mathrm{N}(0,1)$.

(b) Construire, en employant le résultat en (a), un test de la forme $\bar{X} > C$ ayant un risque de première espèce égal à α (où $0 < \alpha < 1$).

(c) Appliquer le test construit en (b), avec $\lambda_0 = 1$ et $\alpha = 0{,}05$, si un échantillon aléatoire particulier a donné le tableau d'effectifs suivant:

k	0	1	2	3	4	5	6	Total
n_k	21	13	8	4	2	1	1	50

(d) Calculer le risque de deuxième espèce du test fait en (c), si la vraie valeur de λ est 1,2.

(e) Quelle est la taille n de l'échantillon à prélever, si l'on veut un risque de deuxième espèce de 0,05 lorsque $\lambda = 1{,}2$?

Question n$^{\text{o}}$ 15

Soit X une variable aléatoire discrète dont la fonction de probabilité est donnée par

$$p_X(k;\theta) = \begin{cases} 1/\theta \text{ si } k = 1, 2, \ldots, \theta & (\theta \in \mathbb{N}) \\ 0 \quad \text{autrement} \end{cases}$$

On a $E[X] = \frac{\theta+1}{2}$ et $\mathrm{VAR}[X] = \frac{\theta^2-1}{12}$.

(a) Un échantillon aléatoire de 120 observations de X a donné le tableau d'effectifs suivant:

k	1	2	3	4	5	6
n_k	14	18	23	19	24	22

Tester l'ajustement du modèle à ces données, avec $\theta = 6$, au seuil de signification $\alpha = 0{,}01$.

(b) On définit $\theta_1 = P[X > 3]$. Utiliser le tableau d'effectifs de la partie (a) pour tester au seuil de signification $\alpha = 0{,}05$

$$H_0\colon \theta_1 = \frac{1}{2} \quad \text{contre} \quad H_1\colon \theta_1 > \frac{1}{2}$$

(c) Calculer la puissance du test effectué en (b), si la vraie valeur du paramètre θ_1 est égale à 0,55.

Question n$^{\text{o}}$ 16

Le tableau qui suit donne la production de 10 machines de type I au cours d'une journée et la production de 10 machines de type II lors d'une autre journée:

	1	2	3	4	5	6	7	8	9	10
Type I	195	200	185	210	210	185	200	180	205	190
Type II	205	195	190	215	225	195	200	205	220	210

Soit X et Y les variables aléatoires qui désignent respectivement la production journalière des machines de type I et celle des machines de type II. On admettra que X et Y présentent approximativement une distribution normale.

(a) Tester, au seuil de signification $\alpha = 0{,}10$, l'égalité des moyennes, si les machines de type II sont une modification de celles de type I et si les 10 mêmes machines ont été utilisées pour recueillir les observations.

(b) Si les machines de type I et celles de type II ne sont pas les mêmes, tester au seuil de signification $\alpha = 0{,}05$

(i) l'égalité des variances;

(ii) l'égalité des moyennes (en tenant compte de (i)).

(c) Supposons que $\sigma_X = 13$.

(i) Tester au seuil de signification $\alpha = 0{,}01$

$$H_0\colon \mu_X = 200 \quad \text{contre} \quad H_1\colon \mu_X < 200$$

(ii) Calculer la probabilité de commettre une erreur de deuxième espèce, si la vraie valeur de la moyenne de X est $\mu_X = 195$.

(iii) Combien d'observations faudrait-il prélever pour que $\beta(\mu_X = 195)$ soit égal à 0,25?

(d) (i) Obtenir un intervalle de confiance à 90 % pour μ_Y.

(ii) Obtenir un intervalle de confiance à 99 % pour σ_Y^2, si la vraie valeur de la moyenne de Y est $\mu_Y = 205$.

Question n° 17

Soit X une variable aléatoire qui présente une distribution exponentielle:

$$f_X(x;\mu) = \frac{1}{\mu} e^{-x/\mu} \quad \text{pour } x > 0$$

On veut tester l'hypothèse

$$H_0\text{: } \mu = 1 \quad \text{contre} \quad H_1\text{: } \mu > 1$$

en utilisant le critère suivant: on rejette H_0 si et seulement si $\bar{X} > C > 0$, où $\bar{X}$ est la moyenne d'un échantillon aléatoire de taille n de X.

(a) Déterminer la constante de rejet C, à l'aide du théorème central limite, afin d'obtenir un test dont le risque de première espèce α est de 0,025, et ce, avec un échantillon de taille n.

(b) Calculer le risque de deuxième espèce pour $n = 25$, $\alpha = 0{,}025$ et $\mu = 2$.

(c) Déterminer la taille n de l'échantillon pour que le risque de deuxième espèce soit de 0,01 avec $\alpha = 0{,}025$ et $\mu = 2$.

Question n° 18

On désire comparer deux machines, A et B. La machine B, d'un coût d'achat plus élevé, devrait produire une proportion plus faible d'articles défectueux que la machine A. Des essais comparatifs sur chaque machine ont donné les résultats suivants:

Machine	Taille de l'échantillon	Nombre de défectueux
A	200	6
B	100	2

(a) Formuler le problème de comparaison des machines sous la forme d'un test d'hypothèses; préciser l'hypothèse nulle, la contre-hypothèse ainsi que la notation employée.

(b) Quelle est la conclusion concernant la comparaison des machines si l'on utilise un risque de première espèce de 0,10?

Remarque. On pourra employer des approximations basées sur la distribution normale pour effectuer les calculs.

(c) Combien d'articles défectueux sur 200 devrait-on avoir avec la machine A, si l'on veut conclure, avec un risque de première espèce de 0,10, que la machine B produit une proportion plus faible d'articles défectueux que la machine A (en supposant comme ci-dessus que, dans le cas de la machine B, on a trouvé deux articles défectueux dans un échantillon de taille 100)?

Question n° 19

On a mesuré la tension de rupture (en kilogrammes par centimètre carré) de 10 types de béton, chacun étant mesuré avec 2 appareils, A et B. Les données sont les suivantes:

	1	2	3	4	5	6	7	8	9	10
A	155	170	160	140	175	150	170	200	120	160
B	160	190	190	210	180	180	240	190	150	180

(a) Tester l'hypothèse que les deux appareils donnent, en moyenne, le même résultat. Utiliser un seuil de signification $\alpha = 0{,}05$.

(b) Supposons que les 20 données sont en fait 20 observations particulières d'un même type de béton. Tester, au seuil de signification $\alpha = 0{,}10$, l'hypothèse que les variances sont égales.

Question n° 20

Pour que l'exploitation d'une certaine mine soit rentable, il faut que le minerai extrait ait une teneur moyenne en métal supérieure à 30 %. On veut être certain à 99 % de ne pas entreprendre l'exploitation d'une mine qui serait non rentable. On suppose que la teneur X (en %) du minerai présente approximativement une distribution normale.

(a) On a évalué cette teneur en 20 points de la mine et on a obtenu les résultats suivants: $\bar{x} = 33$ et $s = 4$. Peut-on entreprendre l'exploitation de la mine? Justifier au moyen d'un test d'hypothèses.

(b) Si la vraie moyenne de X est $\mu = 32$, quelle est la probabilité, en se basant sur les observations recueillies en (a), de ne pas entreprendre l'exploitation de la mine? On supposera que l'écart-type de X est égal à 4.

(c) Supposons, comme en (b), que $\sigma_X = 4$. Combien d'évaluations (indépendantes) de la teneur devrait-on faire, si l'on veut être certain à 95 % d'entreprendre l'exploitation d'une mine pour laquelle $\mu = 33$?

Question n^{o} 21

On désire comparer deux marques de dentifrice. Une étude d'une durée de trois ans a donné les résultats suivants:

Dentifrice	Nombre d'enfants	Nombre total de caries	Écart-type
A	21	416	12,0
B	20	450	10,5

On suppose que le nombre de caries par enfant (en incluant les caries superficielles) présente approximativement une distribution normale. Peut-on conclure qu'une marque est meilleure que l'autre? Utiliser $\alpha = 0{,}05$.

Indication. On a: $F_{0{,}025;20,19} \simeq 2{,}51$ et $t_{0{,}025;39} \simeq 2{,}023$.

Question n^{o} 22

L'entreprise Spectra-Zap fabrique des tubes laser depuis plus de 20 ans. Son meilleur modèle est le HeNe-633, un laser de type hélium-néon (faisceau rouge). Chaque jour, un ingénieur prélève et analyse un échantillon de 20 tubes. Un relevé des 50 dernières journées a permis de constituer le tableau suivant:

X	0	1	2	3	4	5^{+}
n_i	17	20	10	2	1	0

où X désigne le nombre de tubes défectueux dans un échantillon et n_i le nombre d'échantillons.

(a) Soit θ la probabilité qu'un tube soit défectueux. La valeur de l'estimateur à vraisemblance maximale de θ, à partir du tableau de données, est $\theta_{VM} = 0{,}05$. Justifier ce résultat.

(b) Vérifier, à l'aide d'un test d'ajustement du khi-deux de Pearson, si le nombre de tubes défectueux dans un échantillon présente une distribution binomiale de paramètres $n = 20$ et θ. Utiliser $\alpha = 0{,}05$.

Question n^{o} 23

Les articles produits par une certaine machine sont vendus à deux prix différents selon qu'ils sont jugés de première ou de seconde qualité. On estime par expérience que 40 % des articles sont considérés comme de première qualité (et 60 % comme de seconde qualité).

(a) Combien d'observations doit-on recueillir, si l'on veut tester cette hypothèse avec un risque de première espèce égal à 0,05 et un risque de deuxième espèce

égal à 0,10 lorsque le pourcentage véritable d'articles de première qualité est de 35 %?

(b) Dans un échantillon aléatoire particulier de 1000 articles, 365 d'entre eux ont été jugés de première qualité. Que peut-on conclure? Utiliser $\alpha = 0{,}05$.

Question n^{o} 24

Les voitures de marque A et celles de marque B coûtent environ 15.000 $ à l'achat. On désire comparer la valeur de revente de ces voitures après deux ans d'usure, afin de pouvoir choisir plus facilement entre les deux marques. On suppose que le prix de revente présente approximativement une distribution normale et on acceptera que l'écart-type de ce prix de revente est égal à 500 $ dans les deux cas. Un relevé des annonces classées parues dans un journal, durant une semaine, a permis de constituer le tableau suivant:

	1	2	3	4	5	6	7	8	9	10
A	9500	8950	9750	9300	9000	8500	8350	8450	9000	9600
B	9700	10.000	9000	9250	9500	9350	8750	9900	9500	10.500

(a) En se basant sur les sommes demandées par les revendeurs, peut-on affirmer que les voitures de marque B ont une valeur de revente moyenne supérieure à celles de marque A après deux ans d'usure? Utiliser $\alpha = 0{,}05$.

(b) Quel est le risque de deuxième espèce du test effectué en (a), si les voitures de marque B ont en fait une valeur de revente moyenne de 500 $ supérieure à celle des voitures de marque A après deux ans d'usure?

(c) Combien d'observations, au minimum, devrait-on avoir en (b), si l'on veut que cette probabilité soit égale, au plus, à 0,10?

Question n^{o} 25

On désire comparer le degré de satisfaction des résidants de quatre arrondissements de l'île de Montréal. Pour ce faire, on interroge des personnes prises au hasard (et sans remise) dans chacun de ces arrondissements et on leur demande si elles sont: non satisfaites (A), peu satisfaites (B), satisfaites (C) ou très satisfaites (D) de leur arrondissement. On a obtenu les données suivantes:

Arrond. \ Satisfaction	A	B	C	D
I	2	7	33	8
II	3	10	35	12
III	2	5	20	13
IV	3	8	32	7

(a) Peut-on conclure que le degré de satisfaction varie de façon significative d'un arrondissement à l'autre? Utiliser $\alpha = 0{,}05$.

(b) Calculer un intervalle de confiance à 95 % pour la proportion véritable de personnes *non* ou *peu satisfaites* de leur arrondissement (indépendamment de l'arrondissement).

(c) Soit X la variable aléatoire qui désigne le degré de satisfaction des résidants. On pose que

$$X = \begin{cases} 1 \text{ si le résidant est } \textit{non satisfait} \\ 2 \text{ si le résidant est } \textit{peu satisfait} \\ 3 \text{ si le résidant est } \textit{satisfait} \\ 4 \text{ si le résidant est } \textit{très satisfait} \end{cases}$$

Tester, au seuil de signification $\alpha = 0{,}05$, l'ajustement du modèle suivant aux données:

$$p_X(x) = \begin{cases} 0{,}05 \text{ si } x = 1 \\ 0{,}20 \text{ si } x = 2 \\ 0{,}55 \text{ si } x = 3 \\ 0{,}20 \text{ si } x = 4 \end{cases}$$

(d) Tester, avec $\alpha = 0{,}05$,

$$H_0 \colon p_I = p_{II} \quad \text{contre} \quad H_1 \colon p_I > p_{II}$$

où p_I (respectivement p_{II}) est la proportion des résidants de l'arrondissement I (respectivement II) qui sont *satisfaits* ou *très satisfaits* de leur arrondissement.

Question n$^{\text{o}}$ 26

Les données qui suivent sont censées être des observations particulières d'une distribution $\mathrm{N}(0, 1)$:

$$\begin{array}{cccccccccc} 0{,}41 & 1{,}14 & 0{,}33 & 0{,}10 & -1{,}55 & -0{,}01 & 0{,}87 & 1{,}26 & 1{,}47 & -1{,}79 \\ -1{,}46 & -0{,}12 & -0{,}73 & 1{,}92 & -0{,}68 & 0{,}21 & 1{,}20 & -0{,}44 & -0{,}53 & 1{,}11 \end{array}$$

(a) Supposons que les données proviennent effectivement d'une population normale. Tester, au seuil de signification $\alpha = 0{,}05$,

(i) l'hypothèse que la moyenne de la population est nulle;

(ii) l'hypothèse que la variance de la population est égale à 1.

Indication. On a: $t_{0{,}025;39} \simeq 2{,}023$, $\chi^2_{0{,}975;19} \simeq 8{,}91$ et $\chi^2_{0{,}025;19} \simeq 32{,}85$.

(b) Tester, au seuil de signification $\alpha = 0{,}10$, l'hypothèse que les données proviennent vraiment d'une distribution $\mathrm{N}(0,1)$. Utiliser les intervalles

$$(-\infty; -0{,}674)\,, \quad [-0{,}674; 0), \quad [0; 0{,}674) \quad \text{et} \quad [0{,}674; \infty)$$

Question n$^{\text{o}}$ 27

Les entreprises Telmi et Limet fabriquent des puces électroniques pour circuits intégrés. Dans ses brochures, Telmi annonce que la taille moyenne des traits les plus fins de ses puces est de 2,0 dixièmes de micromètre (μm). Par ailleurs, Limet bénéficie d'une nouvelle technologie et prétend produire des puces caractérisées par des traits plus petits que 2,0 dixièmes de μm, en moyenne.

L'entreprise Polypuces, quant à elle, est intéressée par les produits des deux entreprises fabricantes. Afin de pouvoir faire un choix judicieux, Polypuces décide de procéder à l'analyse d'un lot de puces provenant de chacune des deux entreprises.

Soit X_i la variable aléatoire désignant la dimension (en dixièmes de μm) des traits les plus fins pour le lot de l'entreprise i ($i = 1, 2$). On suppose que les variables X_1 et X_2 présentent toutes les deux une distribution approximativement normale et qu'elles sont indépendantes.

Polypuces obtient les résultats suivants:

Lot provenant de Telmi:	Lot provenant de Limet:
- Variable: X_1	- Variable: X_2
- Taille de l'échantillon: $n_1 = 16$	- Taille de l'échantillon: $n_2 = 14$
- Coefficient de variation: 35 %	- $\sum x_{2i} = 21$
- Écart-type de la moyenne: 0,18	- $\sum x_{2i}^2 = 33{,}5$

(a) Calculer les moyennes et variances des deux lots analysés.

(b) Tester, au seuil de signification $\alpha = 0{,}05$, l'hypothèse voulant que Limet produise des puces aux traits plus fins, en moyenne, que Telmi.

Indication. On a: $F_{0{,}025;15,13} \simeq 3{,}05$, $t_{0{,}05;13} \simeq 1{,}771$ et $t_{0{,}05;28} \simeq 1{,}701$.

Question n$^{\text{o}}$ 28

(a) Un fabricant de câbles affirme que ses produits ont une tension de rupture moyenne de 330 kilogrammes ou plus. Peut-on mettre en doute la véracité de cette affirmation, si des expériences faites sur 10 câbles ont permis de mesurer les tensions de rupture suivantes:

251 247 255 305 341
324 329 345 392 289

Utiliser $\alpha = 0{,}10$. Préciser l'hypothèse nulle et la contre-hypothèse.

Indication. On a: $\bar{x} = 307{,}8$ et $s \simeq 47{,}56$. De plus, on supposera que la tension de rupture X présente (approximativement) une distribution normale.

(b) Le fabricant affirme aussi que l'écart-type de la tension de rupture ne dépasse pas 30 kg. Tester cette hypothèse avec un risque de première espèce égal à 0,05.

(c) Quelle aurait été la conclusion en (a), si l'on avait cru le fabricant en prenant $\sigma_X = 30$?

(d) Quelle est la probabilité d'accepter l'affirmation $\sigma_X = 30$ avec le test effectué en (b), lorsque la véritable valeur de l'écart-type de X est $\sigma_X = 45$?

Question n° 29

Un ingénieur veut comparer deux marques, A et B, d'appareils servant à mesurer les émissions, en parties par million, d'oxydes d'azote (NOx) du système d'échappement des voitures. Il considère deux possibilités d'expérimentation (I et II) pour effectuer la comparaison.

Expérience I

Vingt voitures de marques différentes seront utilisées; 10 voitures prises au hasard recevront l'appareil de marque A et les autres recevront l'appareil de marque B.

Expérience II

Les 10 premières voitures de l'expérience I, qui avaient été munies d'un appareil de marque A, seront utilisées une seconde fois, mais avec un appareil de marque B.

On convient de faire les deux expériences, dont les résultats sont présentés dans le tableau suivant:

Expérience I			Expérience II	
Émissions de NOx en ppm			Émissions de NOx en ppm	
Marque A	Marque B	Voiture	Marque A	Marque B
72,1	73,5	1	72,1	74,0
68,2	71,7	2	68,2	68,8
70,9	73,3	3	70,9	71,2
74,3	71,3	4	74,3	74,2
70,7	74,7	5	70,7	71,8
66,6	68,5	6	66,6	66,4
69,5	66,9	7	69,5	69,8
70,8	69,3	8	70,8	71,3
68,8	71,8	9	68,8	69,3
73,3	72,6	10	73,3	73,6

Remarque. Tous les tests seront effectués avec un risque de première espèce (seuil de signification) de 0,05.

(a) Analyser les données de l'expérience I. Préciser la notation et les hypothèses de base utilisées, l'hypothèse nulle et la contre-hypothèse.

(b) Analyser les données de l'expérience II.

(c) Quel serait approximativement le risque de deuxième espèce du test effectué en (b), si l'on avait utilisé 30 voitures et si la différence entre les vraies moyennes des deux appareils était égale à la moitié de l'écart-type de la variable qui désigne la différence entre A et B?

(d) Laquelle des deux expériences semble la plus appropriée pour comparer les appareils? Justifier.

Question n^{o} 30

Le pourcentage moyen de débris produits par une opération de finition d'un métal est censé être de 7 % ou moins. On a pris plusieurs journées au hasard et les pourcentages de débris ont été calculés:

6,5151 7,4970 7,4601 6,3723 8,3257 9,8199 9,5618

On suppose que le pourcentage X de débris présente (approximativement) une distribution $\mathrm{N}(\mu, \sigma^2)$.

(a) En se basant sur les données, peut-on conclure que μ est significativement supérieur à 7 %? Utiliser $\alpha = 0{,}01$.

(b) S'il est important de détecter un rapport $(\mu - 0{,}07)/\sigma$ de 0,5 avec une probabilité d'au moins 95 %, quel est le nombre minimal d'observations que l'on doit recueillir?

(c) Pour un rapport $(\mu - 0{,}07)/\sigma$ de 2, quelle est approximativement la puissance du test effectué en (a)?

Question n° 31

Le diamètre d'une bille en acier a été mesuré par 12 individus en utilisant deux types de compas. Les données suivantes, où par exemple 5 représente 0,265 et 4 représente 0,264, ont été recueillies:

Individu	1	2	3	4	5	6	7	8	9	10	11	12
Compas 1	5	5	6	7	7	5	7	7	5	8	8	5
Compas 2	4	5	4	6	7	8	4	5	5	7	8	9

(a) Y a-t-il une différence significative entre les moyennes des populations dont proviennent les deux échantillons aléatoires particuliers? Utiliser un seuil $\alpha = 0{,}05$ et préciser les hypothèses de base que l'on doit faire pour pouvoir effectuer le test.

Indication. On a: $t_{0{,}025;11} \simeq 2{,}201$.

(b) L'écart-type de la première population est-il significativement supérieur à 0,001? Utiliser $\alpha = 0{,}10$.

Indication. On a: $\chi^2_{0{,}10;11} \simeq 17{,}28$.

(c) Calculer la covariance des données des deux populations.

Question n° 32

Un générateur de nombres aléatoires est censé produire des entiers de 0 à 9 de façon que chaque chiffre ait la même probabilité d'apparaître. Les 10.000 premiers chiffres générés ont permis de constituer le tableau d'effectifs suivant:

0	1	2	3	4	5	6	7	8	9
967	1008	975	1022	1003	989	1001	981	1043	1011

(a) Le générateur fonctionne-t-il bien? Justifier votre réponse en effectuant un test d'ajustement du khi-deux de Pearson au seuil de signification $\alpha = 0{,}05$.

(b) Calculer $F_{10.000}(4{,}5)$, où $F_{10.000}$ est la fonction de répartition de l'échantillon.

Question n° 33

Le fabricant d'un nouvel engrais chimique prétend que, si l'on utilise son engrais, environ 80 % des graines qui ont été semées vont germer. Soit θ la proportion véritable des graines qui germent à l'aide de cet engrais. Dans une expérience réalisée avec 50 graines, on a trouvé que 35 d'entre elles avaient germé. Tester, au seuil de signification $\alpha = 0{,}05$, l'ajustement d'une distribution de Bernoulli de paramètre $\theta = 0{,}8$ aux données.

Question n° 34

La précision de la fabrication en série de boulons est sous contrôle statistique tant que l'écart-type du diamètre des boulons ne dépasse pas 1,5 millimètre. On veut tester, au seuil de signification $\alpha = 0{,}05$, l'hypothèse que cette fabrication est sous contrôle. On suppose que le diamètre X des boulons présente approximativement une distribution $\mathrm{N}(\mu, \sigma^2)$.

(a) Quelle est la conclusion du test, si la somme de 20 mesures du diamètre est de 109,8 et la somme des carrés de 655,6?

Indication. On a: $\chi^2_{0{,}05;19} \simeq 30{,}144$.

(b) Quel est le risque de deuxième espèce du test effectué en (a), si l'écart-type est en fait de 2,2 mm?

(c) Déterminer la taille de l'échantillon pour que le risque de deuxième espèce soit de 0,05 lorsque l'écart-type est de 2,2 mm.

Question n° 35

La résistance X de deux matériaux, A et B, présente approximativement une distribution normale: $X_A \sim \mathrm{N}(\mu_A, \sigma_A^2)$ et $X_B \sim \mathrm{N}(\mu_B, \sigma_B^2)$. De plus, X_A et X_B sont des variables aléatoires indépendantes. Pour tester l'hypothèse

$$H_0\colon \mu_A = \mu_B \quad \text{contre} \quad H_1\colon \mu_A \neq \mu_B$$

on dispose de deux échantillons aléatoires particuliers:

X_A	2,08	3,77	2,46	2,20	2,58	3,35	3,52
X_B	3,13	3,07	3,66	3,51	3,22		

Les calculs donnent:

$$\bar{x}_A = 2{,}85, \quad s_A^2 \simeq 0{,}464; \quad \bar{x}_B = 3{,}32, \quad s_B^2 \simeq 0{,}065$$

(a) Effectuer ce test avec un risque de première espèce α de 0,10.

(b) Effectuer ce test avec un risque de première espèce α de 0,05.

Remarque. En (a) et (b), il faut tester l'égalité des variances avec le même risque de première espèce (10 % et 5 %, respectivement).

Questions à choix multiple

Question n^{o} 1

On a recueilli les observations particulières suivantes d'une variable aléatoire X qui présente une distribution B$(n = 5, p)$:

i	0	1	2	3	4	5
n_i	2	5	7	10	6	2

(A) Estimer la valeur du paramètre p par la méthode de vraisemblance maximale.

(a) 0,081 (b) 0,092 (c) 0,5 (d) 0,519 (e) 0,553

(B) On veut utiliser le test d'ajustement du khi-deux de Pearson pour tester l'hypothèse nulle H_0: $X \sim$ B$(n = 5, p = 1/2)$. Combien de degrés de liberté la statistique utilisée aura-t-elle?

(a) 2 (b) 3 (c) 4 (d) 5 (e) 6

(C) Si l'on a calculé, avec un ensemble de 160 nouvelles observations particulières de X, une valeur de 1,15 pour la statistique utilisée en (b), quelle est la plus grande valeur de α que l'on peut utiliser pour accepter le modèle?

(a) 0,01 (b) 0,05 (c) 0,10 (d) 0,95 (e) 0,99

Question n^{o} 2

La vitesse réelle des voitures de trois marques, lorsque leurs indicateurs de vitesse marquent 100 km/h, a été mesurée pour quatre voitures de chaque marque. Les données sont les suivantes:

	Vitesse réelle			
Marque 1	105	103	108	101
Marque 2	101	97	99	104
Marque 3	102	96	98	103

Soit X_i la vitesse réelle des voitures de marque i. On suppose que

$$X_i \approx \mathrm{N}(\mu + d_i, \sigma^2) \quad \text{pour } i = 1, 2, 3 \quad (\text{avec } \textstyle\sum_{i=1}^{3} d_i = 0)$$

et que les variables aléatoires X_i sont indépendantes. On a: $SC_T \simeq 134{,}92$ et $\hat{\sigma}^2 \simeq 9{,}58$.

(A) Calculer la valeur de la statistique F_0 utilisée pour tester H_0: $d_1 = d_2 = d_3 = 0$.

(a) 2,5 (b) 4,8 (c) 5,1 (d) 12,2 (e) 24,3

(B) Calculer un intervalle de confiance à 95 % pour μ.

(a) 101,42 ± 1,64 (b) 101,42 ± 2,02 (c) 101,42 ± 9,77 (d) 101,42 ± 11,35
(e) 101,42 ± 19,66

Question n$^{\text{o}}$ 3

Soit X le temps (en secondes) requis pour effectuer une certaine opération. On suppose que $X \approx \mathrm{N}\big(\mu, \sigma^2 = 4\big)$. Un échantillon aléatoire particulier de taille $n = 16$ a donné: $\bar{x} = 12$ et $s = 3$. On veut tester

$$H_0\colon \mu = 10 \quad \text{contre} \quad H_1\colon \mu > 10$$

au seuil de signification $\alpha = 0{,}05$.

(A) Quelle est la valeur de la statistique Z_0 utilisée pour effectuer le test?

(a) 0,167 (b) 0,25 (c) 2 (d) 2,67 (e) 4

(B) Quelle est la valeur de β si $\mu = 13$?

(a) $\simeq 0$ (b) 0,01 (c) 0,5 (d) 0,99 (e) $\simeq 1$

(C) Combien d'observations doit-on recueillir pour que $\beta\,(\mu = 13) = 0{,}10$?

(a) 4 (b) 5 (c) 6 (d) 12 (e) 14

Question n$^{\text{o}}$ 4

Soit $X_1 \sim \mathrm{N}\big(\mu_1, \sigma_1^2\big)$ et $X_2 \sim \mathrm{N}\big(\mu_2, \sigma_2^2\big)$ deux variables aléatoires indépendantes. Des échantillons aléatoires particuliers de tailles $n_1 = 25$ et $n_2 = 16$ ont donné les résultats suivants: $\bar{x}_1 = 3$, $\bar{x}_2 = 5$, $s_1 = 2$ et $s_2 = 4$.

(A) On veut tester H_0: $\sigma_1^2 = \sigma_2^2$ contre H_1: $\sigma_1^2 < \sigma_2^2$ au seuil de signification $\alpha = 0{,}01$. Quelle est approximativement la valeur du centile utilisé pour effectuer la comparaison?

(a) 0,304 (b) 0,347 (c) 1,61 (d) 2,88 (e) 3,29

(B) Si l'on accepte que $\sigma_1^2 = \sigma_2^2$, quelle est alors la valeur de la statistique T_0 utilisée pour tester H_0: $\mu_1 - \mu_2 = -1$ contre H_1: $\mu_1 - \mu_2 \neq -1$ au seuil de signification $\alpha = 0{,}05$?

(a) $-3{,}19$ (b) $-2{,}13$ (c) $-1{,}06$ (d) 1,06 (e) 2,13

(C) Quelle est la valeur de β en (b) si $\mu_1 - \mu_2 = -1$?

(a) 0 (b) 0,05 (c) 0,5 (d) 0,95 (e) 1

Question n$^{\text{o}}$ 5

On a relevé le nombre d'accidents de travail dans une certaine usine au cours d'une période d'un an. Le tableau suivant donne les résultats obtenus selon l'équipe de travail et le type de machine utilisée par les employés:

	Type I	Type II
Jour	20	25
Soir	8	7
Nuit	7	8

(A) On veut tester l'hypothèse que, étant donné qu'un accident de travail a eu lieu, la probabilité qu'il se soit produit le jour, sur une machine de type I, est supérieure à 0,5. Calculer la statistique Z_0 utilisée pour effectuer le test.

(a) $-2{,}582$ (b) $-0{,}845$ (c) $-0{,}745$ (d) 0,845 (e) 2,582

(B) Si l'on veut tester l'hypothèse que la proportion des accidents de travail le jour est la même pour les deux types de machines, quelle est la valeur de l'estimateur de cette probabilité commune p?

(a) 4/7 (b) 3/5 (c) 5/8 (d) 2/3 (e) 5/7

(C) Quelle est la valeur de la statistique D^2 utilisée pour tester l'indépendance entre le type de machine et l'équipe de travail (en ce qui concerne les accidents de travail)?

(a) 0,089 (b) 0,214 (c) 0,232 (d) 0,357 (e) 0,358

Question n$^{\text{o}}$ 6 (voir la question n$^{\text{o}}$ 11, page 284)

On suppose qu'une certaine variable aléatoire X présente une distribution de Poisson de paramètre inconnu λ. On a recueilli 500 observations particulières de la variable X, lesquelles sont résumées dans le tableau suivant:

j	0	1	2	3	≥ 4
n_j	260	150	70	20	0

On veut vérifier, à l'aide d'un test d'ajustement du khi-deux de Pearson, si la variable aléatoire X présente effectivement une distribution de Poisson. Quelle est la valeur de la statistique D^2 utilisée pour effectuer le test?

(a) 4,70 (b) 5,70 (c) 6,70 (d) 7,70 (e) 8,70

Question n^o 7

La tension de rupture X (en kilonewtons par mètre carré (kN/m^2)) d'une certaine fibre synthétique présente approximativement une distribution $\mathrm{N}(\mu, 100)$. L'entreprise qui fabrique cette fibre affirme que sa tension de rupture moyenne est d'au moins 240 kN/m^2. Afin de vérifier cette affirmation, on décide d'effectuer un test statistique. Un échantillon aléatoire particulier de taille $n = 18$ a donné $\bar{x} = 237$ kN/m^2.

(A) Quelle est la contre-hypothèse H_1 à considérer pour ce test?

(a) H_1: $\mu < 240$ (b) H_1: $\mu \neq 240$ (c) H_1: $\mu > 240$ (d) H_1: $\mu \neq 237$

(e) H_1: $\mu = 237$

(B) Quelle est la valeur de la statistique Z_0 utilisée pour effectuer le test?

(a) $-2{,}031$ (b) $-1{,}273$ (c) $-1{,}031$ (d) $-0{,}127$ (e) 1,031

(C) Calculer la valeur de β si $\mu = 235$, $n = 18$ et $\alpha = 0{,}05$.

(a) 0,32 (b) 0,68 (c) 0,90 (d) 0,95 (e) $\simeq 1$

(D) Quel est le nombre minimal d'observations que l'on doit recueillir pour que la probabilité de rejeter l'affirmation de l'entreprise soit d'au moins 0,975 lorsque $\mu = 236$ et $\alpha = 0{,}05$?

(a) 9 (b) 81 (c) 82 (d) 324 (e) 328

Question n^o 8

Dans le but de comparer la perte de chaleur dans un tuyau en acier, d'une part, et dans un tuyau en verre, d'autre part, on considère sept paires de tuyaux (un en acier et un en verre). Les tuyaux d'une paire ont le même diamètre, mais les diamètres sont différents d'une paire à l'autre. On fait circuler de l'eau dans les tuyaux à la même température initiale et l'on mesure la perte de chaleur (en degrés Celsius) sur une distance de 50 mètres. Les résultats obtenus sont les suivants:

N^o de la paire	1	2	3	4	5	6	7
Tuyau en acier	4,6	3,7	4,2	1,9	4,8	6,1	4,7
Tuyau en verre	2,5	4,2	2,0	1,8	2,7	3,2	3,0

On suppose que la perte de chaleur présente approximativement une distribution normale.

(A) Quelle est la valeur de la statistique T_0 utilisée pour tester l'hypothèse selon laquelle la perte de chaleur est en moyenne plus grande dans les tuyaux en acier que dans les tuyaux en verre?

(a) $-4{,}38$ (b) $-2{,}62$ (c) 2,62 (d) 3,24 (e) 4,38

(B) Combien de degrés de liberté possède la statistique du test d'hypothèses en (A)?

(a) 3 (b) 4 (c) 5 (d) 6 (e) 12

Question n^{o} 9

On dispose des résultats suivants calculés à partir d'un échantillon aléatoire particulier de 200 articles classés selon leur état (qualité) et la période de la semaine où ils ont été fabriqués:

	Lundi et mardi	Mercredi	Jeudi et vendredi
Bons	70	32	68
Défectueux	10	8	12

(A) Quelle est la valeur de la statistique Z_0 utilisée pour tester l'hypothèse selon laquelle au plus 22 % des articles produits le mercredi sont défectueux?

(a) $-0{,}305$ (b) $-0{,}049$ (c) 0,049 (d) 0,305 (e) 1,049

(B) Pour tester l'indépendance entre la période de fabrication et la qualité d'un article, quelle est la valeur de la statistique D^2 utilisée pour effectuer le test?

(a) 0,20 (b) 0,96 (c) 1,18 (d) 2,12 (e) 3,24

(C) Combien de degrés de liberté possède la statistique du test d'indépendance en (B)?

(a) 2 (b) 3 (c) 4 (d) 5 (e) 6

7

Régression linéaire simple

On a vu au chapitre précédent comment tester l'hypothèse que les données que l'on a recueillies proviennent d'une distribution particulière, comme la distribution normale. Pour ce faire, on peut se servir des tests d'ajustement de Pearson, de Shapiro et Wilk (pour la normalité) et de Kolmogorov et Smirnov.

Lorsque la variable aléatoire qui nous intéresse est une fonction d'une variable déterministe (que l'on peut contrôler), la *régression* peut nous aider à trouver la forme de la relation entre les variables aléatoire et déterministe. Compte tenu du fait que, dans tous les domaines du génie et des sciences naturelles, les étudiants et les chercheurs effectuent des expériences scientifiques au cours desquelles des données sont recueillies, le sujet de la régression est très important.

Dans ce chapitre, nous allons traiter en détail le problème de base, soit celui de trouver la meilleure relation linéaire entre la variable aléatoire et la variable déterministe. Nous allons aussi considérer brièvement le cas où la relation entre les deux variables est supposée non linéaire.

Nous ne traiterons pas ici le cas où l'on cherche une relation entre une variable aléatoire et deux ou plusieurs variables déterministes. La *régression multiple* exige l'utilisation de logiciels statistiques lors de la résolution de problèmes pratiques, pour l'inversion de matrices, par exemple. Cependant, une fois la régression linéaire simple bien comprise, il est possible de généraliser la théorie sans trop de difficultés.

7.1 Le modèle

Soit Y une variable aléatoire et x une variable déterministe (c'est-à-dire non aléatoire). On dispose d'un échantillon aléatoire $(x_1, Y_1), \ldots, (x_n, Y_n)$ et on désire

trouver une relation mathématique qui exprime Y en fonction de x. La variable x est appelée variable **indépendante** et Y variable **dépendante** ou variable **réponse**.

Dans le cas de la *régression linéaire simple*, le modèle que l'on propose est de la forme

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{7.1}$$

où ϵ est un terme d'erreur. On suppose que chaque observation Y_i de Y satisfait à l'équation

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{7.2}$$

où $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ pour $i = 1, \ldots, n$ et les ϵ_i sont des variables aléatoires *indépendantes*. Notons que l'on tient pour acquis que la variance des ϵ_i est la même pour tout i.

On peut donc écrire que

$$Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2) \quad \text{pour } i = 1, \ldots, n \tag{7.3}$$

car β_0 et β_1 sont des paramètres (c'est-à-dire des constantes).

Remarque. Le modèle est appelé modèle de régression linéaire simple, car il n'y a qu'une seule variable indépendante et le modèle est linéaire par rapport aux *paramètres*. Le modèle

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{7.4}$$

est linéaire par rapport aux paramètres et polynomial de degré deux, tandis que

$$Y_i = \beta_0 e^{\beta_1 x_i} + \epsilon_i \tag{7.5}$$

est un modèle non linéaire par rapport aux paramètres.

Les meilleurs estimateurs des paramètres β_0 et β_1, c'est-à-dire les estimateurs non biaisés et à variance minimale de β_0 et β_1, sont obtenus en utilisant la *méthode des moindres carrés*. On définit la somme

$$SC = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{7.6}$$

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1 par la méthode des moindres carrés sont les valeurs de β_0 et β_1 qui minimisent la somme SC. On pose:

$$\frac{\partial SC}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{7.7}$$

et

$$\frac{\partial SC}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(Y_i - \beta_0 - \beta_1 x_i) = 0 \tag{7.8}$$

La solution de ces deux équations, appelées **équations normales**, est

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \tag{7.9}$$

Pour prédire la valeur de Y lorsque $x = x_i$, on utilise alors l'équation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{7.10}$$

Remarques.
(i) Dans le cas où l'on suppose que les erreurs ϵ_i sont indépendantes et présentent toutes une distribution $\mathrm{N}(0, \sigma^2)$, comme ci-dessus, on trouve que les estimateurs donnés par la méthode des moindres carrés et ceux obtenus par la méthode de vraisemblance maximale sont les mêmes.

(ii) Les ϵ_i sont aussi appelés **résidus** (théoriques); la quantité ϵ_i représente la différence entre l'observation Y_i et la valeur calculée à partir du modèle proposé $Y_i = \beta_0 + \beta_1 x_i$.

(iii) L'équation de prédiction ne devrait (théoriquement) être utilisée que pour des valeurs de x dans l'intervalle $[x_{(1)}, x_{(n)}]$, où

$$x_{(1)} := \min\{x_1, \ldots, x_n\} \quad \text{et} \quad x_{(n)} := \max\{x_1, \ldots, x_n\} \tag{7.11}$$

(iv) Même si l'on sait que $Y = 0$ lorsque $x = 0$, on ne pose pas que $\beta_0 = 0$; on obtient généralement de meilleurs résultats en laissant β_0 dans le modèle. Cependant, si l'on désire effectivement proposer le **modèle de régression passant par l'origine**:

$$Y = \beta x + \epsilon \tag{7.12}$$

où $\epsilon \sim \mathrm{N}(0, \sigma^2)$, on trouve facilement que l'estimateur du paramètre β obtenu par la méthode des moindres carrés, à partir d'un échantillon aléatoire $(x_1, Y_1), \ldots, (x_n, Y_n)$, est

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \tag{7.13}$$

Exemple 7.1.1 (voir la référence [19]). On veut déterminer de quelle façon la force de tension d'un certain alliage dépend du pourcentage de zinc qu'il contient. On a les données suivantes:

% de zinc	4,7	4,8	4,9	5,0	5,1
Force de tension	1,2	1,4	1,5	1,5	1,7

On obtient le graphique de la figure 7.1.

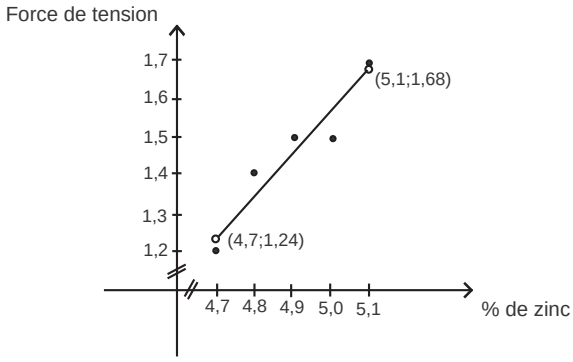

Fig. 7.1. Graphique dans l'exemple 7.1.1

On propose le modèle de régression linéaire simple:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

où x est le pourcentage de zinc et Y la force de tension.

Remarque. Il est important de bien repérer la variable aléatoire et la variable déterministe dans le problème.

On trouve que

$$\bar{x} = 4{,}9, \quad \bar{y} = 1{,}46, \quad \sum_{i=1}^{5} x_i^2 = 120{,}15 \quad \text{et} \quad \sum_{i=1}^{5} x_i y_i = 35{,}88$$

Alors on a:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{5} x_i y_i - 5\bar{x}\bar{y}}{\sum_{i=1}^{5} x_i^2 - 5\bar{x}^2} = \frac{35{,}88 - 5(4{,}9)(1{,}46)}{120{,}15 - 5(4{,}9)^2} = 1{,}1$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1{,}46 - (1{,}1)(4{,}9) = -3{,}93$$

Donc, l'équation de prédiction est donnée par

$$\hat{y} = -3{,}93 + 1{,}1x$$

♢

7.2 Tests d'hypothèses

Étant donné que les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des combinaisons linéaires de variables aléatoires normales indépendantes, ils présentent aussi une distribution normale. De plus, on peut montrer que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs *non biaisés* de β_0 et β_1, et que

$$\text{VAR}[\hat{\beta}_0] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SC_x} \right) \quad \text{et} \quad \text{VAR}[\hat{\beta}_1] = \frac{\sigma^2}{SC_x} \tag{7.14}$$

où

$$SC_x := \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{7.15}$$

Remarque. Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ ne sont généralement *pas* indépendants. En fait, on peut montrer que

$$\text{COV}[\hat{\beta}_0, \hat{\beta}_1] = -\frac{\sigma^2 \bar{x}}{SC_x} \tag{7.16}$$

Puisque deux variables aléatoires normales sont indépendantes *si et seulement si* leur covariance (ou leur coefficient de corrélation) égale zéro (voir la page 163), on peut affirmer que $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des variables aléatoires indépendantes si et seulement si $\bar{x} = 0$. Cependant, soit

$$\hat{\beta}_0' = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{7.17}$$

On trouve que $\mathrm{COV}[\hat{\beta}_0', \hat{\beta}_1] = 0$, de sorte que $\hat{\beta}_0'$ et $\hat{\beta}_1$ *sont* des estimateurs indépendants.

Ensuite, on définit la **somme des carrés des erreurs** (ou des résidus)

$$SC_E = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{7.18}$$

On peut montrer que

$$E[SC_E] = (n-2)\sigma^2 \tag{7.19}$$

Il s'ensuit qu'un estimateur non biaisé de la variance σ^2 est donné par (le **carré moyen des erreurs**)

$$\hat{\sigma}^2 = CM_E := \frac{SC_E}{n-2} \tag{7.20}$$

(a) Pour tester l'hypothèse nulle

$$H_0\colon \beta_0 = \beta_{00} \tag{7.21}$$

on utilise la statistique

$$T_0 := \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{CM_E\left(\frac{1}{n} + \frac{\bar{x}^2}{SC_x}\right)}} \overset{H_0}{\sim} t_{n-2} \tag{7.22}$$

On rejette alors H_0 au seuil de signification α si et seulement si

$$\begin{cases} |T_0| > t_{\alpha/2,n-2} & \text{si } H_1\colon \beta_0 \neq \beta_{00} \\ T_0 > t_{\alpha,n-2} & \text{si } H_1\colon \beta_0 > \beta_{00} \\ T_0 < -t_{\alpha,n-2} & \text{si } H_1\colon \beta_0 < \beta_{00} \end{cases} \tag{7.23}$$

(b) De même, pour faire le test de

$$H_0\colon \beta_1 = \beta_{10} \tag{7.24}$$

on utilise la statistique

$$T_0 := \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{CM_E/SC_x}} \overset{H_0}{\sim} t_{n-2} \tag{7.25}$$

et on rejette alors H_0 au seuil de signification α si et seulement si

$$\begin{cases} |T_0| > t_{\alpha/2,n-2} & \text{si } H_1\text{: } \beta_1 \neq \beta_{10} \\ T_0 > t_{\alpha,n-2} & \text{si } H_1\text{: } \beta_1 > \beta_{10} \\ T_0 < -t_{\alpha,n-2} & \text{si } H_1\text{: } \beta_1 < \beta_{10} \end{cases} \tag{7.26}$$

Cas particulier où $\beta_{10} = 0$: test de la signification globale de la régression

Le test de

$$H_0\text{: } \beta_1 = 0 \quad \text{contre} \quad H_1\text{: } \beta_1 \neq 0 \tag{7.27}$$

est très important, car si l'on ne peut pas rejeter l'hypothèse nulle H_0, alors on doit conclure qu'il n'y a pas de relation *linéaire* entre x et Y (ou, à tout le moins, on ne peut pas conclure qu'il y a une relation linéaire *significative* entre x et Y). Pour effectuer le test, on peut procéder comme ci-dessus. En pratique, on effectue plutôt, de façon équivalente, une **analyse de la variance**. Notons d'abord que, lorsqu'on choisit $\beta_{10} = 0$, on a:

$$T_0^2 = \frac{\hat{\beta}_1^2}{CM_E/SC_x} \tag{7.28}$$

On peut montrer que

$$\hat{\beta}_1^2 SC_x = SC_R := \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{7.29}$$

On dit que la somme des carrés SC_R est la **somme des carrés due à la régression**. On peut aussi montrer que les sommes SC_R et SC_E sont des variables aléatoires *indépendantes* qui présentent des distributions du khi-deux à 1 et $n-2$ degrés de liberté, respectivement. Il s'ensuit que

$$F_0 := T_0^2 = \frac{SC_R}{CM_E} \overset{H_0}{\sim} F_{1,n-2} \tag{7.30}$$

Remarques.
(i) En fait, par définition, le carré d'une distribution de Student à n degrés de liberté est une distribution de Fisher à 1 et n degrés de liberté. Donc, on pouvait écrire directement que $T_0^2 \sim F_{1,n-2}$, lorsque l'hypothèse nulle H_0 est vraie.
(ii) Puisque SC_R possède un seul degré de liberté, le carré moyen CM_R et la somme des carrés SC_R sont égaux.

Finalement, on peut montrer que

$$E[SC_R] = \sigma^2 + \beta_1^2 SC_x \overset{H_0}{=} \sigma^2 \tag{7.31}$$

Puisque $E[CM_E] = \sigma^2$, plus la valeur de la statistique F_0 est grande, plus la probabilité que l'hypothèse nulle H_0 soit vraie est faible. On rejette H_0 au seuil de signification α si et seulement si

$$F_0 > F_{\alpha,1,n-2} \tag{7.32}$$

La façon de procéder pour tester la signification globale de la régression peut être résumée en utilisant un tableau d'analyse de la variance, dans lequel la **somme des carrés totale** SC_T est définie par

$$SC_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2 \tag{7.33}$$

Analyse de la variance

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens	F_0
Régression	SC_R	1	CM_R	$\frac{CM_R}{CM_E}$
Erreur	SC_E	$n-2$	CM_E	
Totale	SC_T	$n-1$		

En pratique, on calcule d'abord la somme des carrés SC_T; ensuite, on a: $SC_R \equiv \hat{\beta}_1^2 SC_x = \hat{\beta}_1 SC_{xY}$, où

$$SC_{xY} := \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} \tag{7.34}$$

On peut alors compléter le tableau d'analyse de la variance.

Exemple 7.2.1. On a trouvé dans l'exemple 7.1.1 que

$$\hat{y} = -3{,}93 + 1{,}1x$$

On veut maintenant tester

$$H_0\colon \beta_1 = 0 \quad \text{contre} \quad H_1\colon \beta_1 \neq 0$$

On a:

$$SC_T = \sum_{i=1}^{5} y_i^2 - 5\bar{y}^2 = 10{,}79 - 5(1{,}46)^2 = 0{,}132$$

et

$$SC_R = \hat{\beta}_1 SC_{xY} = (1,1)\left(\sum_{i=1}^{5} x_i y_i - 5\bar{x}\bar{y}\right) = (1{,}1)(35{,}88 - 35{,}77) = 0{,}121$$

Ainsi, $SC_E = SC_T - SC_R = 0{,}011$ et

$$f_0 = \frac{CM_R}{CM_E} = \frac{0{,}121}{0{,}011/3} = 33$$

On trouve dans le tableau des $F_{0{,}05;n_1,n_2}$, à la page 519, que $F_{0{,}05;1,3} \simeq 10{,}1$. Puisque $33 > 10{,}1$, on peut rejeter H_0 au seuil de signification $\alpha = 0{,}05$. Notons qu'en fait on a: $F_{0{,}05;1,3} = t^2_{0{,}025,3}$. Donc, on peut aussi se servir du tableau 5.2, à la page 236, pour obtenir les valeurs critiques $F_{0{,}05;1,n-2}$. ♢

7.3 Intervalles et ellipses de confiance

En utilisant les résultats suivants:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{SC_x}} \sim t_{n-2} \quad \text{et} \quad \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\bar{x}^2}{SC_x}}} \sim t_{n-2} \tag{7.35}$$

où $\hat{\sigma} = \sqrt{CM_E} = \sqrt{SC_E/(n-2)}$, on peut calculer des intervalles de confiance pour les paramètres β_0 et β_1. On trouve que les intervalles de confiance bilatéraux à $100(1-\alpha)$ % sont donnés (sous forme compacte) respectivement par

$$\hat{\beta}_0 \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SC_x}} \tag{7.36}$$

et

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{SC_x}} \tag{7.37}$$

De même, on peut montrer qu'un intervalle de confiance à $100(1-\alpha)$ % pour la valeur moyenne de la variable aléatoire Y lorsque $x = \xi \in [x_{(1)}, x_{(n)}]$, c'est-à-dire pour $\beta_0 + \beta_1\xi$, est

$$\hat{\beta}_0 + \hat{\beta}_1\xi \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(\xi - \bar{x})^2}{SC_x}} \tag{7.38}$$

On voit que la longueur de l'intervalle est minimale lorsque $\xi = \bar{x}$.

Maintenant, la formule précédente est valable lorsqu'on désire construire un intervalle de confiance pour $\beta_0 + \beta_1\xi$, basé sur les n observations recueillies. Dans le cas où l'on désire plutôt un intervalle de confiance pour une *nouvelle* observation (c'est-à-dire pour une *prédiction*) de Y à $x = \xi$, la variance augmente de σ^2 et la formule devient

$$\hat{\beta}_0 + \hat{\beta}_1\xi \pm t_{\alpha/2,n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\xi - \bar{x})^2}{SC_x}} \tag{7.39}$$

Remarque. Si l'on veut un intervalle de confiance pour la *moyenne* de m nouvelles observations de Y à $x = \xi$, il suffit de remplacer le terme 1 (ajouté devant $\frac{1}{n}$) par $\frac{1}{m}$ dans la formule précédente.

Finalement, on peut montrer que l'intérieur de l'*ellipse* définie par

$$n(\beta_0 - \hat{\beta}_0)^2 + 2n\bar{x}(\beta_0 - \hat{\beta}_0)(\beta_1 - \hat{\beta}_1) + (\beta_1 - \hat{\beta}_1)^2 \sum_{i=1}^{n} x_i^2 = 2\hat{\sigma}^2 F_{\alpha,2,n-2} \tag{7.40}$$

est un région de confiance à $100(1-\alpha)$ % pour le couple (β_0, β_1).

Exemple 7.3.1. Un intervalle de confiance bilatéral à 95 % pour une nouvelle observation de Y lorsque $x = \bar{x} = 4{,}9$, dans l'exemple 7.1.1, est donné par

$$-3{,}93 + 1{,}1(4{,}9) \pm t_{0{,}025;3}\hat{\sigma}\sqrt{1 + \frac{1}{5} + 0} \simeq 1{,}46 \pm 0{,}21$$

car $\hat{\sigma} = \sqrt{CM_E} = \sqrt{0{,}011/3}$ (voir l'exemple 7.2.1) et $t_{0{,}025;3} \overset{\text{tab. } 5.2}{\simeq} 3{,}182$. $\diamondsuit$

7.4 Le coefficient de détermination

Définition 7.4.1. *Le* **coefficient de détermination**, R^2, *est obtenu en divisant la somme des carrés SC_R par la somme des carrés SC_T:*

$$R^2 = \frac{SC_R}{SC_T} \tag{7.41}$$

Le coefficient R^2, aussi appelé **coefficient de corrélation au carré**, permet de mesurer l'ajustement du modèle aux données; si tous les points se trouvent sur la droite de régression, alors on a: $R^2 = 1$. En général, R^2 prend une valeur dans l'intervalle $[0, 1]$ et nous donne le pourcentage de la variation totale SC_T qui est expliquée par le modèle de régression. Il s'ensuit que

$$1 - R^2 = \frac{SC_E}{SC_T} \tag{7.42}$$

est la proportion de SC_T qui n'est *pas* expliquée par le modèle.

Remarques.
(i) L'expression *coefficient de corrélation* est en fait abusive, car x n'est pas une variable aléatoire. En effet, un coefficient de corrélation n'est défini que pour des variables aléatoires X et Y.

(ii) La quantité $R = \sqrt{R^2}$ est appelée *indice d'ajustement*. On l'utilise souvent pour mesurer la qualité du modèle de régression. Cependant, il est important de savoir que, si la valeur de R est grande, cela ne signifie pas *nécessairement* que le modèle de régression linéaire (simple) est le bon modèle. De même, si la valeur de R est petite, il ne faut pas conclure que le modèle doit être rejeté. L'indice d'ajustement est en fait une mesure de l'amélioration de l'ajustement obtenue en ajoutant le terme $\beta_1 x$ dans le modèle, comparativement au modèle de régression $Y = \beta_0 + \epsilon$.

Exemple 7.4.1. En utilisant les résultats de l'exemple 7.2.1, on trouve que

$$R^2 = \frac{0{,}121}{0{,}132} \simeq 0{,}917$$

pour les données de l'exemple 7.1.1. ♢

7.5 L'analyse des résidus

On a déjà appelé les quantités $\epsilon_i = Y_i - \beta_0 - \beta_1 x_i$ *résidus* (théoriques). On définit les **résidus de l'échantillon** par

$$e_i = y_i - \hat{y}_i \tag{7.43}$$

L'analyse des résidus, qui se fait généralement à l'aide d'un logiciel statistique, permet de vérifier, en particulier, la validité de l'hypothèse que l'on a faite au début de ce chapitre, à savoir que les termes d'erreur ϵ_i présentent tous une distribution normale (de moyenne nulle et de variance commune σ^2). Si l'hypothèse est vraie, alors les quantités

$$z_i := \frac{e_i}{\hat{\sigma}} \tag{7.44}$$

appelées **résidus standardisés**, devraient être des observations particulières d'une variable aléatoire Z qui présente approximativement une distribution normale centrée réduite. Puisque

$$P[-2 < Z < 2] \simeq 0{,}95 \quad \text{et} \quad P[-3 < Z < 3] \simeq 0{,}997 \tag{7.45}$$

il ne devrait pas y avoir plus de 5 % (environ) des z_i tels que $|z_i| \geq 2$ et presque aucun z_i tel que $|z_i| \geq 3$.

Remarque. Si la taille de l'échantillon aléatoire est petite, alors il faut utiliser la distribution de Student. C'est-à-dire que l'on doit comparer les z_i aux valeurs prises par une distribution t_{n-2}.

Exemple 7.5.1. L'équation de régression obtenue avec les données fournies dans l'exemple 7.1.1 est

$$\hat{y} = -3{,}93 + 1{,}1x$$

et on a: $\hat{\sigma} = \sqrt{0{,}011/3}$. En utilisant ces résultats, on peut construire le tableau suivant:

x_i	y_i	$\hat{y}_i$	e_i	z_i
4,7	1,2	1,24	−0,04	−0,67
4,8	1,4	1,35	0,05	0,83
4,9	1,5	1,46	0,04	0,67
5,0	1,5	1,57	−0,07	−1,17
5,1	1,7	1,68	0,02	0,33

On voit que tous les résidus standardisés sont petits; alors l'hypothèse de normalité des variables aléatoires ϵ_i semble raisonnable.

Remarque. Puisque $n = 5$, on devrait comparer les résidus aux valeurs d'une distribution t_3. Or, on a:

$$P[-3{,}182 \leq T \leq 3{,}182] \simeq 0{,}05$$

si $T \sim t_3$. Donc, on pouvait en fait accepter que les résidus soient situés dans l'intervalle $[-3{,}182; 3{,}182]$. Cependant, la taille de l'échantillon aléatoire particulier est un peu petite ici pour faire l'analyse des résidus. ♢

L'analyse des résidus permet aussi de vérifier les autres hypothèses de base que l'on fait en régression linéaire simple:

(a) le modèle est de la forme $Y = \beta_0 + \beta_1 x$;

(b) les erreurs ϵ_i sont indépendantes et possèdent la même variance σ^2.

Pour ce faire, on trace les graphiques des résidus e_i en fonction des x_i et des $\hat{y}_i$. Dans tous les cas, les points dans ces graphiques devraient former une bande uniforme, comme dans la figure 7.2.

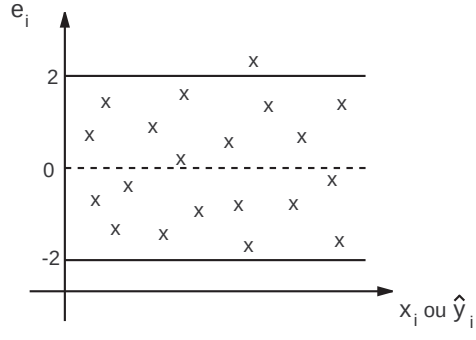

Fig. 7.2. Résidus formant une bande uniforme

La figure 7.3, à la page 394, illustre des cas où l'on peut conclure que

(i) le modèle n'est pas linéaire en x (graphiques (a) et (b));

(ii) la variance des erreurs n'est pas constante (graphiques (c) et (d));

(iii) les deux à la fois (graphiques (e) et (f)).

Remarque. S'il existe au moins une valeur de la variable indépendante x pour laquelle on possède au moins deux observations de la variable dépendante Y (et

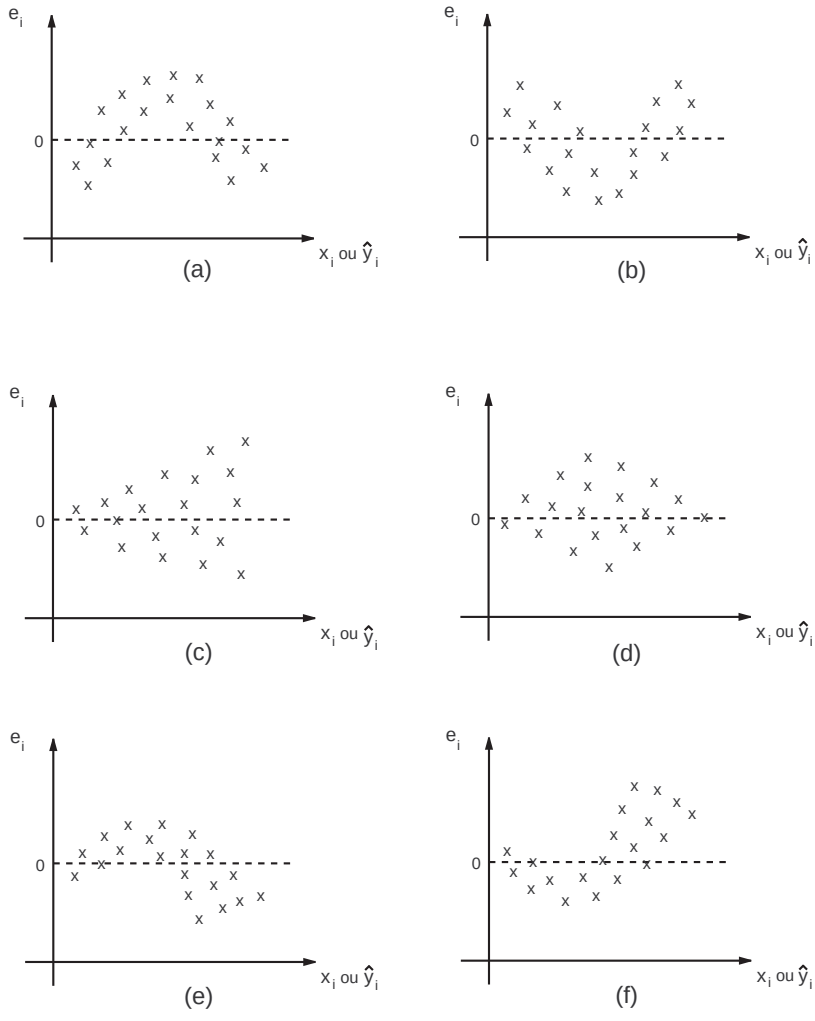

Fig. 7.3. Résidus indiquant au moins une hypothèse non vérifiée

s'il y a au moins trois valeurs différentes de x en tout), alors on peut effectuer le test suivant de l'ajustement du modèle aux données: soit $Y_{i,1}, \ldots, Y_{i,n_i}$ les n_i observations de la variable aléatoire Y lorsque $x = x_i$, pour $i = 1, \ldots, k$. On définit la **somme des carrés attribuable entièrement à l'erreur**:

$$SC_{EE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2 \tag{7.46}$$

où

$$\bar{Y}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} \tag{7.47}$$

et on pose:

$$SC_E = SC_{EE} + SC_M \tag{7.48}$$

où SC_M est la **somme des carrés due au manque d'ajustement du modèle**. Ensuite, on considère la statistique

$$F_0 := \frac{SC_M/(k-2)}{SC_{EE}/(n-k)} \tag{7.49}$$

où $n := \sum_{i=1}^{k} n_i$ est le nombre total d'observations dans l'échantillon aléatoire. Le test consiste à rejeter, au seuil de signification α, l'hypothèse que le modèle est adéquat pour les données lorsque $F_0 > F_{\alpha,k-2,n-k}$.

7.6 Régression curviligne

Parfois, un modèle de régression non linéaire peut être ramené à un modèle de régression linéaire en transformant la variable Y ou la variable x (ou les deux à la fois). Par exemple, si l'on propose le modèle (sans terme d'erreur)

$$Y = \beta_0 e^{\beta_1 x} \tag{7.50}$$

alors il suffit de prendre le logarithme (naturel) de chaque côté de l'équation, puis de poser que $Y' = \ln Y$ et $\beta_0' = \ln \beta_0$. On obtient ainsi le modèle

$$Y' = \beta_0' + \beta_1 x \tag{7.51}$$

De même, le modèle

$$Y = \frac{1}{\exp(\beta_0 + \beta_1 x)} \tag{7.52}$$

devient

$$Y' = \beta_0 + \beta_1 x \tag{7.53}$$

si l'on définit $Y' = -\ln Y$.

Remarques.
(i) Si l'on suppose en (7.51) que

$$Y_i' = \beta_0' + \beta_1 x_i + \epsilon_i \tag{7.54}$$

où $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ pour $i = 1, \ldots, n$, alors cela implique que le modèle original est

$$Y_i = \beta_0 e^{\beta_1 x_i + \epsilon_i} \tag{7.55}$$

Un terme d'erreur de la forme e^{ϵ_i} peut ne pas être réaliste dans plusieurs situations. Par contre, supposer que le terme d'erreur présente une distribution *log-normale* peut être intéressant dans certains cas.

(ii) Les modèles non linéaires que l'on peut linéariser sont dits *intrinsèquement* linéaires.

(iii) On se rend compte de la non-linéarité de la relation entre x et Y à partir du graphique des y_i en fonction des x_i, ou de celui des e_i en fonction des $\hat{y}_i$.

Exemple 7.6.1. On dispose du tableau de données suivant:

x	1	2	3	7
Y	13,0	21,9	29,8	30,4
	11,8	24,7	24,1	35,7

où x est le temps (en jours) après la prise d'un béton de ciment et Y la tension (en kilogrammes par centimètre carré) du béton. On propose le modèle de régression curviligne

$$Y = \beta_0 e^{-\beta_1 x}$$

(a) Transformer les variables pour obtenir un modèle linéaire.

(b) Estimer les paramètres β_0 et β_1 par la méthode des moindres carrés.

(c) Calculer un intervalle de prédiction à 95 % pour Y lorsque $x = 5$ et 10.

Remarque. Lorsqu'il y a plus d'une valeur de Y pour une même valeur de x, on devrait spécifier dans l'énoncé si toutes les observations sont indépendantes (en fait, on devrait toujours spécifier dans l'énoncé que les observations sont indépendantes). Ici, on suppose que l'on dispose effectivement de *huit* valeurs prises par des observations indépendantes.

Solution. (a) On a:

$$Y = \beta_0 e^{-\beta_1 x} \iff \ln Y = \ln \beta_0 - \beta_1 x$$

Il suffit donc de poser que $Y' = \ln Y$, $\beta_0' = \ln \beta_0$ et $\beta_1' = -\beta_1$ pour obtenir

$$Y' = \beta'_0 + \beta'_1 x$$

Remarque. On aurait aussi pu poser que $x' = -x$, plutôt que $\beta'_1 = -\beta_1$.

(b) Il faut d'abord transformer les données; on obtient:

x	1	2	3	7
$\simeq Y'$	2,565	3,086	3,395	3,414
	2,468	3,207	3,182	3,575

On calcule ensuite $\hat{\beta}'_0$ et $\hat{\beta}'_1$ en utilisant les formules de la section 7.1. On trouve que

$$\hat{\beta}'_0 \simeq 2{,}69 \quad \text{et} \quad \hat{\beta}'_1 \simeq 0{,}13$$

Il s'ensuit que

$$\hat{\beta}_0 = e^{\hat{\beta}'_0} \simeq 14{,}74 \quad \text{et} \quad \hat{\beta}_1 = -\hat{\beta}'_1 \simeq -0{,}13$$

(c) L'intervalle de prédiction à 95 % pour Y', lorsque $x = \xi$, est donné par

$$\hat{\beta}'_0 + \hat{\beta}'_1 \xi \pm \underbrace{t_{0{,}025;6}}_{2{,}447} \hat{\sigma} \sqrt{1 + \frac{1}{8} + \frac{(\xi - \bar{x})^2}{SC_x}}$$

On trouve que $\bar{x} = 3{,}25$, $SC_x = 41{,}5$ et (avec les données transformées) $SC_T \simeq 1{,}114$. Il s'ensuit que $SC_E = SC_T - (\hat{\beta}'_1)^2 \, SC_x \simeq 0{,}413$ et alors

$$\hat{\sigma} \simeq \left(\frac{0{,}413}{6} \right)^{1/2} \simeq 0{,}262$$

Ainsi, on obtient que

$$\xi = 5 \implies 3{,}34 \pm 0{,}70 \quad \text{et} \quad \xi = 10 \implies 3{,}99 \pm 0{,}96$$

Finalement, en fonction de la variable Y, les intervalles de prédiction deviennent

$$Y \mid \{x = 5\} \in [14{,}0; 56{,}8] \quad \text{et} \quad Y \mid \{x = 10\} \in [20{,}7; 141{,}2]$$

Remarques.
(i) On voit, en regardant les données, que Y augmente lorsque x augmente; il est donc logique que $\hat{\beta}_1$ soit négatif.

(ii) Puisque β_1 est un paramètre réel, le modèle proposé est équivalent au modèle $Y = \beta_0 e^{\beta_1 x}$.

(iii) Si l'on avait proposé le modèle $Y = \beta_0 + \beta_1 \ln x$, par exemple, alors il aurait suffi de poser que $x' = \ln x$ pour obtenir le modèle de régression linéaire simple. De plus, si l'on suppose que

$$Y_i = \beta_0 + \beta_1 x_i' + \epsilon_i$$

où $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ pour $i = 1, \ldots, n$, alors on a:

$$Y_i = \beta_0 + \beta_1 \ln x_i + \epsilon_i$$

C'est-à-dire que, si l'on transforme uniquement la variable indépendante x, alors cela ne change pas le terme d'erreur ϵ_i dans le modèle. ♢

7.7 Corrélation

Lorsque x n'est pas une variable déterministe, mais plutôt une variable aléatoire X, on suppose que $(X_1, Y_1), \ldots, (X_n, Y_n)$ est un échantillon aléatoire de taille n du vecteur aléatoire (X, Y). Si l'on suppose, en outre, que X présente une distribution $\mathrm{N}(\mu_X, \sigma_X^2)$ et Y une distribution $\mathrm{N}(\mu_Y, \sigma_Y^2)$, alors on peut montrer que la moyenne et la variance de Y, étant donné que $X = x$, sont données par

$$E[Y \mid X = x] = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \tag{7.56}$$

et

$$\mathrm{VAR}[Y \mid X = x] = \sigma_Y^2 (1 - \rho_{X,Y}^2) \tag{7.57}$$

Il s'ensuit que l'on peut écrire que

$$E[Y \mid X = x] = \beta_0 + \beta_1 x \tag{7.58}$$

où

$$\beta_0 := \mu_Y - \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} \mu_X \quad \text{et} \quad \beta_1 := \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} \tag{7.59}$$

On peut montrer que les estimateurs à vraisemblance maximale des paramètres β_0 et β_1 sont donnés par

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n Y_i (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{7.60}$$

et

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{7.61}$$

Notons que ces estimateurs ont la même forme que ceux obtenus par la méthode des moindres carrés à la section 7.1.

Maintenant, l'estimateur $\hat{\rho}_{X,Y}$ du coefficient de corrélation théorique $\rho_{X,Y}$ est le **coefficient de corrélation empirique** (ou **de l'échantillon**) $R_{X,Y}$, défini par (voir le chapitre 5, page 212)

$$R_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}} := \frac{SC_{XY}}{\sqrt{SC_X SC_Y}} \tag{7.62}$$

Remarque. On trouve que le carré du coefficient de corrélation empirique, $r_{X,Y}$, d'un échantillon aléatoire particulier du couple (X, Y) est égal au coefficient de détermination: $r_{X,Y}^2 = R^2$. Cependant, le coefficient de détermination a été défini pour une variable déterministe x (et une variable aléatoire Y). Comme nous l'avons déjà mentionné, il n'est donc pas rigoureux d'affirmer que le coefficient de détermination est le carré du coefficient de corrélation empirique des variables x et Y.

Test d'hypothèses

Pour tester l'hypothèse

$$H_0\colon \rho_{X,Y} = 0 \quad \text{contre} \quad H_1\colon \rho_{X,Y} \neq 0 \tag{7.63}$$

on utilise la statistique

$$T_0 := \sqrt{n-2}\frac{R_{X,Y}}{\sqrt{1 - R_{X,Y}^2}} \overset{H_0}{\sim} t_{n-2} \tag{7.64}$$

On rejette H_0 au seuil de signification α si et seulement si $|T_0| > t_{\alpha/2,n-2}$.

Remarques.
(i) Puisque $SC_T = SC_Y$ et $\hat{\beta}_1 = SC_{XY}/SC_X$, on peut écrire que

$$R_{X,Y} = \hat{\beta}_1 \left(\frac{SC_X}{SC_T}\right)^{1/2} \tag{7.65}$$

Alors tester $H_0\colon \rho_{X,Y} = 0$ est équivalent, au point de vue mathématique, à tester $H_0\colon \beta_1 = 0$.

(ii) Pour tester l'hypothèse plus générale H_0: $\rho_{X,Y} = \rho_0$, on peut utiliser la statistique

$$Z_0 := \frac{\sqrt{n-3}}{2}\left[\ln\left(\frac{1+R_{X,Y}}{1-R_{X,Y}}\right) - \ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right] \tag{7.66}$$

qui présente approximativement une distribution $\mathrm{N}(0,1)$ si l'hypothèse H_0 est vraie et si la taille n de l'échantillon aléatoire est assez grande. On rejette H_0 au seuil de signification α si et seulement si $|Z_0| > z_{\alpha/2}$.

Exemple 7.7.1. Les données suivantes sont les températures maximales et minimales (en degrés Fahrenheit) enregistrées au cours d'une semaine d'hiver dans une ville américaine:

	Dim.	Lun.	Mar.	Mer.	Jeu.	Ven.	Sam.
Maximum	6	11	14	12	5	−2	−9
Minimum	−22	−17	−15	−9	−24	−29	−35

Soit Y la température maximale et X la température minimale. On calcule $\bar{x} \simeq -21{,}57$ et $\bar{y} \simeq 5{,}29$. De plus,

$$SC_X = \sum_{i=1}^{7} x_i^2 - 7\bar{x}^2 \simeq 463{,}71$$

$$SC_Y = \sum_{i=1}^{7} y_i^2 - 7\bar{y}^2 \simeq 411{,}43$$

$$SC_{X,Y} = \sum_{i=1}^{7} x_i y_i - 7\bar{x}\bar{y} \simeq 414{,}14$$

Il s'ensuit que

$$\hat{\beta}_1 = \frac{SC_{X,Y}}{SC_X} \simeq 0{,}89 \quad \text{et} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \simeq 24{,}55$$

de sorte que l'on estime $\mu_{Y|x} := E[Y \mid X = x]$ par

$$\hat{\mu}_{Y|x} \simeq 24{,}55 + 0{,}89x$$

Remarque. On pourrait aussi calculer

$$\hat{\mu}_{X|y} \simeq -26{,}89 + 1{,}01y$$

Le coefficient de corrélation empirique $r_{X,Y}$ est donné par

$$r_{X,Y} = \frac{SC_{X,Y}}{\sqrt{SC_X SC_Y}} \simeq 0{,}9482$$

Puisque $R^2 = r_{X,Y}^2 \simeq 0{,}8990$, on dit que le modèle explique environ 89,9 % de la variation dans les données.

Finalement, pour tester

$$H_0\colon \rho_{X,Y} = 0 \quad \text{contre} \quad H_1\colon \rho_{X,Y} \neq 0$$

on calcule la statistique

$$t_0 := \sqrt{n-2}\frac{r_{X,Y}}{\sqrt{1 - r_{X,Y}^2}} \simeq 6{,}67$$

Étant donné que $t_{0{,}025;5} \overset{\text{tab. 5.2}}{\simeq} 2{,}571$, on peut rejeter H_0 au seuil de signification $\alpha = 0{,}05$ et conclure que le coefficient de corrélation $\rho_{X,Y}$ n'est pas nul. ♢

7.8 Exercices du chapitre 7

Exercices résolus

Question n$^{\text{o}}$ 1 (voir la référence [25])

En 1957, l'ingénieur industriel néerlandais J. R. DeJong a proposé le modèle suivant pour le temps requis pour accomplir une tâche manuelle simple, en fonction du nombre de fois que la tâche a été effectuée:

$$T = \gamma\beta^{-k}$$

où T est le temps (en secondes), k est le nombre de fois que la tâche a été effectuée, et β et γ sont des paramètres qui dépendent de la tâche et de l'individu. On dispose des données suivantes:

T	22,4	21,3	19,7	15,6	15,2	13,9	13,7
k	0	1	2	3	4	5	6

(a) Transformer le modèle proposé en un modèle linéaire et estimer les paramètres β et γ par la méthode des moindres carrés.

(b) On suppose que

$$T = \gamma\beta^{-k}e^{\epsilon}, \quad \text{où } \epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

(i) Estimer le paramètre σ.

(ii) Tester l'hypothèse H_0: $\beta = 1$ contre H_1: $\beta \neq 1$ avec un risque de première espèce $\alpha = 0{,}05$.

(iii) Calculer un intervalle de confiance approximatif à 95 % pour γ.

Solution. (a) On a:

$$T = \gamma\beta^{-k} \iff \ln T = \ln\gamma - (\ln\beta)\, k$$

On définit

$$Y = \ln T, \quad \beta_0 = \ln\gamma, \quad \beta_1 = -\ln\beta \quad \text{et} \quad x = k$$

Les données deviennent:

$(\simeq)\ Y$	3,11	3,06	2,98	2,75	2,72	2,63	2,62
x	0	1	2	3	4	5	6

On trouve que

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{7} x_i y_i - 7\bar{x}\bar{y}}{\sum_{i=1}^{7} x_i^2 - 7\bar{x}^2} \simeq \frac{57{,}02 - (7)(3)(2{,}84)}{91 - 7\,(3)^2} \simeq -0{,}094$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \simeq 2{,}84 - (-0{,}094)\,(3) \simeq 3{,}12$$

Alors on peut écrire que

$$\hat{\gamma} = e^{\hat{\beta}_0} \simeq 22{,}6 \quad \text{et} \quad \hat{\beta} = e^{-\hat{\beta}_1} \simeq 1{,}1$$

(b) (i) On calcule

$$SC_T = \sum_{i=1}^{7} y_i^2 - 7\bar{y}^2 \simeq 0{,}26$$

et

$$SC_R = \hat{\beta}_1^2 \left(\sum_{i=1}^{7} x_i^2 - 7\bar{x}^2 \right) \simeq 0{,}24$$

Alors on a:

$$\hat{\sigma} = \left(\frac{SC_T - SC_R}{7-2} \right)^{1/2} \simeq 0{,}06$$

(ii) On a: $\beta = 1 \Leftrightarrow \beta_1 = 0$. On rejette H_0: $\beta_1 = 0$ si et seulement si

$$f_0 := \frac{SC_R}{\hat{\sigma}^2} \simeq 73 > F_{0{,}05;1,7-2} \overset{\text{p. } 519}{\simeq} 6{,}61$$

Donc, on rejette H_0.

(iii) L'intervalle de confiance approximatif à 95 % pour β_0 est

$$3{,}12 \pm \underbrace{t_{0{,}025;7-2}}_{2{,}571} \hat{\sigma} \left(\frac{1}{7} + \frac{9}{91 - 7(3)^2} \right)^{1/2} \simeq 3{,}12 \pm 0{,}105$$

Alors celui pour γ est $[20{,}4; 25{,}2]$ (environ).

Question n$^{\text{o}}$ 2

Un lundi, un éleveur de poissons ensemence un grand bassin avec des poissons d'élevage. Chaque lundi subséquent, il attrape au hasard 500 poissons, les marque d'un chiffre (différent d'une semaine à l'autre), puis les remet à l'eau. Au bout de quelques heures, il lance un filet, compte le nombre total de poissons qu'il a pris et le nombre de ceux marqués du chiffre de la semaine, avant de remettre les poissons à l'eau. Il obtient le tableau chronologique suivant du nombre de poissons marqués et du nombre total de poissons attrapés:

Semaine	1	2	3	4	5	6	7	8
# marqués	17	19	18	23	24	26	27	29
# attrapés	1004	1011	1008	1015	1003	1017	1003	1013

Remarque. La précision des calculs est importante pour ce problème.

(a) Soit P la proportion de poissons marqués dans le bassin. L'éleveur considère le modèle suivant à compter de la quatrième semaine:

$$P = \alpha e^{\beta t} \quad \text{pour } t \in [4, 8]$$

Estimer les paramètres α et β par la méthode des moindres carrés.

(b) En extrapolant le modèle proposé, à combien estime-t-on le nombre de poissons dans le bassin au bout de la 40$^{\text{e}}$ semaine?

(c) En supposant que les poissons restent comestibles et qu'ils gagnent 7 % en poids chaque semaine, à partir de la quatrième semaine jusqu'à leur maturité au bout de 50 semaines, l'éleveur a-t-il intérêt à récolter la huitième semaine ou à attendre la maturité?

Remarque. La valeur des poissons est proportionnelle au poids multiplié par le nombre.

Solution. (a) On a:

$$P = \alpha e^{\beta t} \iff \ln P = \ln \alpha + \beta t \iff Y = \beta_0 + \beta_1 x$$

où

$$Y := \ln P, \quad \beta_0 := \ln \alpha, \quad \beta_1 := \beta \quad \text{et} \quad x := t$$

On estime β_0 et β_1 à partir des données suivantes:

x	4	5	6	7	8
$\simeq Y$	$-3{,}7815$	$-3{,}73270$	$-3{,}66652$	$-3{,}61491$	$-3{,}55338$

où $-3{,}7815 \simeq \ln\left(\frac{23}{1015}\right)$, etc. On a:

$$\begin{array}{ccc} n = 5 & \sum_{i=1}^5 x_i = 30 & \sum_{i=1}^5 x_i^2 = 190 \\ \sum_{i=1}^5 y_i \simeq -18{,}35465 & \sum_{i=1}^5 y_i^2 \simeq 67{,}41295 & \sum_{i=1}^5 x_i y_i \simeq -109{,}54258 \end{array}$$

Alors

$$\hat{\beta}_1 = \frac{\sum_{i=1}^5 x_i y_i - 5\bar{x}\bar{y}}{\sum_{i=1}^5 x_i^2 - 5\bar{x}^2} \simeq \frac{-109{,}54258 - (6)\,(-18{,}35465)}{190 - 5\,(6)^2} \simeq 0{,}05853$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \simeq -4{,}02213$$

Il s'ensuit que les valeurs des estimateurs de α et β par la méthode des moindres carrés sont

$$\hat{\alpha} = e^{\hat{\beta}_0} \simeq 0{,}01791 \quad \text{et} \quad \hat{\beta} = \hat{\beta}_1 \simeq 0{,}05853$$

(b) Le nombre $\hat{N}(t)$ de poissons au bout de t semaines, estimé par le modèle, est donné par

$$\hat{N}(t) = \frac{500}{\hat{P}(t)} = \frac{500}{\hat{\alpha}} e^{-\hat{\beta} t} \quad \text{pour } t \geq 4$$

Alors

$$\hat{N}(40) \simeq \frac{500}{0{,}01791} e^{-(0{,}05853)(40)} \simeq 2686$$

(c) La valeur des poissons au bout de t semaines, estimée par le modèle, est

$$V(t) = K(1,07)^{t-4} \frac{500}{\hat{\alpha}} e^{-\hat{\beta}t} \quad \text{pour } t \geq 4$$

où K est une constante de proportionnalité. On a:

$$\frac{V(50)}{V(8)} = (1{,}07)^{50-8} e^{-\hat{\beta}(50-8)} \simeq 1{,}47 > 1$$

Donc, il est préférable d'attendre la maturité.

Question n^{o} 3

On a recueilli les données suivantes:

	1	2	3	4	5	6	7	8	9	10
x	12	12	10	10	10	3	2	20	20	30
Y	20	40	30	80	50	50	90	30	40	40

On trouve que

$$\sum_{i=1}^{10} x_i = 129, \qquad \sum_{i=1}^{10} y_i = 470, \quad \sum_{i=1}^{10} x_i^2 = 2301, \quad \sum_{i=1}^{10} y_i^2 = 26.500$$
$$\sum_{i=1}^{10} x_i y_i = 5250, \quad SC_T = 4410$$

On considère le modèle

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

(a) Quel pourcentage de la variation totale des valeurs de Y peut être expliqué par une relation linéaire entre les variables Y et x?

(b) Peut-on rejeter l'hypothèse H_0: $\beta_1 = 0$ (contre H_1: $\beta_1 \neq 0$) si l'on prend $\alpha = 0{,}05$?

Solution. (a) Le coefficient R^2 est donné par $\frac{SC_R}{SC_T}$, où $SC_T = 4410$ et

$$SC_R = \hat{\beta}_1^2 \left(\sum_{i=1}^{10} x_i^2 - 10\,\bar{x}^2 \right) = (636{,}9)\,(\hat{\beta}_1)^2 \simeq 1037{,}8$$

car

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - 10\,\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - 10\,\bar{x}^2} \simeq -1{,}2765$$

Donc, on a:

$$R^2 \simeq 0{,}2353$$

Remarque. On peut calculer $r_{x,Y}$ avec la calculatrice, par exemple, et alors $R^2 = r_{x,Y}^2$.

(b) On rejette H_0, lorsque $\alpha = 0{,}05$, si et seulement si

$$f_0 := \frac{SC_R}{\hat{\sigma}^2} \simeq \frac{1037{,}8}{(4410 - 1037{,}8)\,/\,(10-2)} \simeq 2{,}46 > F_{0{,}05;1,8} \overset{\text{p. } 519}{\simeq} 5{,}32$$

Donc, on ne peut pas rejeter H_0 si l'on prend $\alpha = 0{,}05$.

Question n$^{\text{o}}$ 4

On croit que la force de tension Y d'une certaine fibre synthétique est reliée au pourcentage x_1 de coton dans la fibre et au temps de séchage x_2 de la fibre. Un test de 10 pièces de cette fibre, produites sous différentes conditions, a donné les résultats suivants:

Y	213	220	216	225	235	218	239	243	233	240
x_1	13	15	14	18	19	20	22	17	16	18
x_2	2,1	2,3	2,2	2,5	3,2	2,4	3,4	4,1	4,0	4,3

Remarque. On a: $\sum_{i=1}^{10} x_{1i} = 172$, $\sum_{i=1}^{10} x_{1i}^2 = 3028$, $\sum_{i=1}^{10} x_{2i} = 30{,}5$ et $\sum_{i=1}^{10} x_{2i}^2 = 99{,}65$.

(a) On propose d'abord les modèles de régression linéaire simple:

$$Y = \beta_0 + \beta_i x_i + \varepsilon, \quad \text{où } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

pour $i = 1, 2$. Tester, au seuil de signification $\alpha = 0{,}05$, chacune des hypothèses nulles H_0: $\beta_i = 0$ contre H_1: $\beta_i \neq 0$, pour $i = 1, 2$.

(b) Si l'on se base sur les quantités R^2 et $\hat{\sigma}^2$, laquelle des deux variables, x_1 ou x_2, semble la plus importante pour expliquer la force de tension de la fibre?

Remarque. Plus la valeur de $\hat{\sigma}^2$ est petite et plus le modèle semble adéquat pour les données.

(c) On considère aussi le modèle

$$Y = \exp(\beta_1 x_1 + \varepsilon), \quad \text{où } \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

Utiliser les données recueillies pour estimer le paramètre β_1 par la méthode des moindres carrés.

Solution. (a) On calcule d'abord $SC_T = 1105{,}6$. Dans le cas du modèle de régression linéaire simple avec la variable x_1, on a:

$$SC_R = \hat{\beta}_1^2 \left(\sum_{i=1}^{10} x_{1i}^2 - 10\bar{x}_1^2 \right) = 69{,}6\, \hat{\beta}_1^2$$

On trouve que $\hat{\beta}_1 \simeq 2{,}221$. Il s'ensuit que $SC_R \simeq 343{,}33$ et

$$SC_E \simeq 762{,}27 \implies \hat{\sigma}^2 \simeq \frac{762{,}27}{8} \simeq 95{,}3 \implies f_0 \simeq \frac{343{,}33}{95{,}3} \simeq 3{,}6$$

On rejette H_0: $\beta_1 = 0$ au seuil de signification $\alpha = 0{,}05$ si et seulement si $f_0 > F_{0{,}05;1,10-2} \overset{\text{p. 519}}{\simeq} 5{,}32$. Donc, on ne peut pas rejeter H_0.

Lorsqu'on considère le modèle avec la variable x_2, on obtient que $SC_R \simeq 942{,}11$, de sorte que

$$SC_E \simeq 163{,}49 \implies \hat{\sigma}^2 \simeq 20{,}4 \implies f_0 \simeq \frac{942{,}11}{20{,}4} \simeq 46{,}1$$

Donc, ici on peut rejeter H_0: $\beta_2 = 0$ au seuil de signification $\alpha = 0{,}05$, car $46{,}1 > 5{,}32$.

(b) Dans le cas de la variable x_1, on a: $\hat{\sigma}^2 \overset{\text{(a)}}{\simeq} 95{,}3$ et

$$R^2 = \frac{SC_R}{SC_T} \overset{\text{(a)}}{\simeq} 0{,}31$$

tandis que, pour la variable x_2, on obtient que $\hat{\sigma}^2 \overset{\text{(a)}}{\simeq} 20{,}4$ et

$$R^2 \overset{\text{(a)}}{\simeq} \frac{942{,}11}{1105{,}6} \simeq 0{,}85$$

Étant donné que R^2 est plus grand et que la valeur de $\hat{\sigma}^2$ est plus petite avec la variable x_2, celle-ci est la plus importante des deux pour expliquer la force de tension.

(c) On a:

$$\ln Y = \beta_1 x_1 + \varepsilon$$

On pose:

$$SC\left(\beta_1\right) = \sum_{i=1}^{10} \left(\ln y_i - \beta_1 x_{1i}\right)^2$$

Alors on a:

$$\frac{d}{d\beta_1} SC\left(\beta_1\right) = 0 \iff 2\sum_{i=1}^{10} \left(\ln y_i - \beta_1 x_{1i}\right)\left(-x_{1i}\right) = 0 \implies \hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_{1i} \ln y_i}{\sum_{i=1}^{10} x_{1i}^2}$$

On trouve que

$$\sum_{i=1}^{10} x_{1i} \ln y_i \simeq 934{,}5 \quad \text{et} \quad \sum_{i=1}^{10} x_{1i}^2 = 3028$$

Il s'ensuit que

$$\hat{\beta}_1 \simeq \frac{934{,}5}{3028} \simeq 0{,}31$$

Question n$^{\text{o}}$ 5

On étudie la production Y (en mètres cubes par seconde) d'un processus, en fonction de la température x (en degrés Celsius). Pour des températures allant (de 100 en 100) de 100 à 600 $^{\text{o}}$C, la production a augmenté de 49 à 68. On donne:

$$\sum_{i=1}^{6} y_i = 369, \quad \sum_{i=1}^{6} y_i^2 = 22.947 \quad \text{et} \quad \sum_{i=1}^{6} x_i y_i = 135.400$$

(a) On propose le modèle de régression linéaire simple:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \text{où } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

(i) Estimer les paramètres β_0 et β_1 par la méthode des moindres carrés.

(ii) Calculer le pourcentage de variation expliqué par le modèle.

(iii) Tester, au seuil de signification $\alpha = 0{,}01$, l'hypothèse H_0: $\beta_1 = 0$ contre H_1: $\beta_1 \neq 0$.

(b) On propose aussi le modèle de régression curviligne suivant:

$$Y = \alpha_0 + \alpha_1\sqrt{x} + \varepsilon, \quad \text{où } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

On trouve alors que $\sum_{i=1}^{6} \sqrt{x_i} y_i \simeq 6847{,}934$.

(i) Calculer les estimations ponctuelles des paramètres α_0 et α_1 par la méthode des moindres carrés.

(ii) Calculer le pourcentage de variation expliqué par le modèle.

Solution. (a) (i) On a: $\sum_{i=1}^{6} x_i = 2100$, $\bar{x} = 350$ et $\sum_{i=1}^{6} x_i^2 = 910.000$; alors

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{6} x_i y_i - 6\bar{x}\bar{y}}{\sum_{i=1}^{6} x_i^2 - 6\bar{x}^2} = \frac{135.400 - 6(350)(61{,}5)}{910.000 - 6\,(350)^2} \\ &= \frac{6250}{175.000} = \frac{1}{28} \simeq 0{,}0357\end{aligned}$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 61{,}5 - \frac{350}{28} = 49$$

(ii) On calcule

$$\begin{aligned}R^2 = \frac{SC_R}{SC_T} &= \frac{\hat{\beta}_1^2\left(\sum_{i=1}^{6} x_i^2 - 6\bar{x}^2\right)}{\sum_{i=1}^{6} y_i^2 - 6\bar{y}^2} \\ &= \left(\frac{1}{28}\right)^2 \frac{175.000}{22.947 - 6\,(61{,}5)^2} \simeq 0{,}8805\end{aligned}$$

(iii) On rejette H_0: $\beta_1 = 0$ si et seulement si

$$f_0 := \frac{SC_R}{SC_E} > F_{0,01;1,4} = t_{0,005;4}^2 \overset{\text{tab. } 5.2}{\simeq} (4{,}604)^2 \simeq 21{,}20$$

Puisque

$$SC_R \overset{\text{(ii)}}{=} \frac{175.000}{(28)^2} \simeq 223{,}21 \quad \text{et} \quad SC_T \overset{\text{(ii)}}{=} 22.947 - 6(61{,}5)^2 = 253{,}5$$

on peut écrire que

$$f_0 \simeq \frac{223{,}21}{(253{,}5 - 223{,}21)/4} \simeq 29{,}48$$

Donc, on peut rejeter H_0 au seuil de signification $\alpha = 0{,}01$.

(b) (i) On trouve d'abord que $\sum_{i=1}^{6} \sqrt{x_i} \simeq 108{,}3182$. Il s'ensuit, avec $u := \sqrt{x}$, que

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{6} u_i y_i - 6\bar{u}\bar{y}}{\sum_{i=1}^{6} u_i^2 - 6\bar{u}^2} \simeq \frac{6847{,}934 - 6\,(108{,}3182/6)\,(61{,}5)}{2100 - 6\left(\frac{108{,}3182}{6}\right)^2}$$
$$\simeq \frac{186{,}365}{144{,}528} \simeq 1{,}289$$

et

$$\hat{\alpha}_0 = \bar{y} - \hat{\alpha}_1 \bar{u} \simeq 61{,}5 - (1{,}289)\left(\frac{108{,}3182}{6}\right) \simeq 38{,}230$$

(ii) On peut écrire que

$$SC_R \simeq \frac{(186{,}365)^2}{144{,}528} \simeq 240{,}313$$

Étant donné que $SC_T = 253{,}5$, le pourcentage de variation expliqué par le modèle est donné par

$$R^2 = \frac{SC_R}{SC_T} \simeq 94{,}8\ \%$$

Question n$^{\text{o}}$ 6

Une entreprise de camionnage veut déterminer la relation entre l'âge d'un camion et le nombre de jours par année qu'il passe en réparation. Pour ce faire, elle a pris six camions au hasard et obtenu les données suivantes:

x	8	1	6	3	5	2
Y	9	16	1	4	0	10

où x est l'âge du camion (en années) et Y le nombre de jours qu'il a passés en réparation au cours d'une période d'une année.

(a) Sans faire de calculs, expliquer pourquoi, parmi les trois modèles qui suivent, le modèle (2) semble le plus approprié:

$$(1) \qquad Y = \beta_0 + \beta_1 x + \varepsilon$$
$$(2) \qquad Y = \beta_0 + \beta_1 x^2 + \varepsilon$$
$$(3) \qquad Y = \beta_0 + \beta_1 e^{-x} + \varepsilon$$

où $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$.

(b) Si l'on se base uniquement sur le coefficient de détermination R^2, lequel, entre les modèles (1) et (3) ci-dessus, est le meilleur modèle?

Solution. (a) Si l'on fait le graphique des y_i en fonction des x_i, on voit que celui-ci a l'allure d'une parabole; donc, le modèle (2) est le plus approprié.

(b) On trouve (avec une calculatrice) que

$$\text{Modèle (1)} : r \simeq -0{,}5034 \implies R^2 \simeq 0{,}2534$$

$$\text{Modèle (3)} : r \simeq +0{,}8391 \implies R^2 \simeq 0{,}7041$$

Donc, le modèle (3) est le meilleur, car $0{,}7041 > 0{,}2534$.

Remarque. Il faut transformer les données au préalable dans le cas du modèle (3).

Question n$^{\text{o}}$ 7

Des chercheurs pensent qu'il existe une relation entre la résistance Y (en dizaines de kilogrammes) d'une certaine pièce métallique et le temps x (en minutes) alloué pour son refroidissement après sa fabrication. Ils mesurent la résistance de 10 pièces refroidies pendant des temps différents et obtiennent les résultats qui sont résumés ci-dessous:

$$\bar{x} = 40{,}9, \;\; \bar{y} = 74{,}4, \;\; \sum_{i=1}^{10} x_i^2 = 17.077, \;\; \sum_{i=1}^{10} y_i^2 = 55.504, \;\; \sum_{i=1}^{10} x_i y_i = 30.436$$

On propose le modèle de régression linéaire simple:

$$Y = \beta_0 + \beta_1 x + \varepsilon, \quad \text{où } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

(a) Estimer les paramètres β_0 et β_1 par la méthode des moindres carrés.

(b) Estimer le paramètre σ.

(c) Tester l'hypothèse H_0: $\beta_1 = 0$ contre H_1: $\beta_1 \neq 0$. Utiliser $\alpha = 0{,}05$.

Solution. (a) On a:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - 10\bar{x}\bar{y}}{\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2} = \frac{30.436 - (10)\,(40{,}9)\,(74{,}4)}{17.077 - (10)\,(40{,}9)^2} = \frac{6{,}4}{348{,}9} \simeq 0{,}0183$$

et

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \simeq 74{,}4 - (0{,}0183)\,(40{,}9) \simeq 73{,}6498$$

(b) On calcule

$$SC_T = \sum_{i=1}^{10} y_i^2 - 10\bar{y}^2 = 55.504 - (10)\,(74{,}4)^2 = 150{,}4$$

et

$$SC_R = \hat{\beta}_1^2 \left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2 \right) \overset{\text{(a)}}{=} \hat{\beta}_1^2\,(348{,}9) \overset{\text{(a)}}{\simeq} 0{,}117$$

Alors on peut écrire que

$$\hat{\sigma} = \left(\frac{SC_T - SC_R}{10 - 2} \right)^{1/2} \simeq 4{,}33$$

(c) On rejette H_0: $\beta_1 = 0$ si et seulement si

$$f_0 := \frac{SC_R}{\hat{\sigma}^2} \overset{\text{(b)}}{\simeq} \frac{0{,}117}{(4{,}33)^2} \simeq 0{,}0062 > F_{0{,}05;1,10-2}$$

Puisque $F_{0{,}05;1,8} \overset{\text{p. 519}}{\simeq} 5{,}32$, on ne peut pas rejeter H_0 si $\alpha = 0{,}05$.

Exercices

Question n° 1

Des calculs préliminaires effectués sur une série d'observations particulières (x_i, y_i), $i = 1, \ldots, 100$, d'une variable aléatoire Y et d'une variable déterministe x ont donné:

$$\sum_{i=1}^{100} x_i = 500, \qquad \sum_{i=1}^{100} y_i = 1000, \qquad \sum_{i=1}^{100} x_i^2 = 24.775$$

$$\sum_{i=1}^{100} (y_i - \bar{y})^2 = 39.600, \qquad r_{x,Y} = 0,9$$

où $r_{x,Y}$ est le "coefficient de corrélation" de l'échantillon, défini par

$$r_{x,Y} = \frac{\sum_{i=1}^{100} (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{100} (x_i - \bar{x})^2 \sum_{i=1}^{100} (y_i - \bar{y})^2 \right]^{1/2}}$$

Un premier examen graphique des données suggère un modèle de régression linéaire simple:

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

(a) Calculer $\sum_{i=1}^{100} (x_i - \bar{x})(y_i - \bar{y})$.

(b) Estimer tous les paramètres du modèle.

(c) Effectuer une analyse de la variance et tester l'hypothèse nulle H_0: $\beta_1 = 0$ au seuil de signification $\alpha = 0{,}05$.

Indication. On a: $F_{0{,}05;1,98} \simeq 3{,}94$.

Question n$^{\text{o}}$ 2

On nous donne le tableau suivant de la progression d'une épidémie au cours des semaines:

t	10	12	15	20	23	25	27	30
N	40	35	30	25	20	20	15	10

où t désigne le nombre de semaines écoulées depuis le début de l'épidémie et N est le nombre (en dizaines) de personnes atteintes. On cherche à prédire la valeur de N au bout de la 40$^{\text{e}}$ semaine. On propose le modèle suivant:

$$N = \beta_0 + \beta_1 t + \varepsilon, \quad \text{où } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

On trouve que

$$n = 8, \qquad \sum_{i=1}^{8} t_i = 162, \qquad \sum_{i=1}^{8} t_i^2 = 3652$$

$$\sum_{i=1}^{8} n_i = 195, \qquad \sum_{i=1}^{8} n_i^2 = 5475, \qquad \sum_{i=1}^{8} t_i n_i = 3435$$

(a) Estimer tous les paramètres du modèle et calculer le tableau d'analyse de la variance.

(b) Donner la prédiction de la valeur de N au bout de la 40$^{\text{e}}$ semaine.

(c) *Rappel.* On définit un intervalle de confiance unilatéral avec borne supérieure à $100(1-\alpha)$ % pour un paramètre θ par la donnée d'une statistique T_S telle que

$$P\left[\theta < T_S\right] = 1 - \alpha$$

Calculer un intervalle de confiance unilatéral avec borne supérieure à 99 % pour la valeur de N au bout de la 40$^{\text{e}}$ semaine.

Question n$^{\text{o}}$ 3

On s'intéresse à la relation entre la vitesse x (en kilomètres à l'heure) atteinte en 10 secondes par 15 voitures de luxe et la distance de freinage Y (en mètres) jusqu'à l'arrêt. Les données suivantes ont été recueillies:

x	104,2	106,1	105,6	106,3	101,7	104,4	102,0	103,8
Y	39,8	40,4	39,9	40,8	33,7	39,5	33,0	37,0

x	104,0	101,5	101,9	100,6	104,9	106,2	103,1
Y	37,0	33,2	33,9	29,9	39,5	40,6	35,1

(a) Estimer les paramètres du modèle

$$Y = \beta_0 + \beta_1\, x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}(0, \sigma^2)$$

(b) Calculer le tableau d'analyse de la variance et tester l'hypothèse nulle H_0: $\beta_1 = 0$ au seuil de signification $\alpha = 0{,}05$.

(c) Calculer des intervalles de confiance à 95 % pour les coefficients β_0 et β_1.

(d) Effectuer une analyse des résidus pour s'assurer que les hypothèses de base sont vérifiées.

(e) Calculer un intervalle de prédiction pour Y lorsque $x = 100$, 103 et 106.

Question n$^{\text{o}}$ 4

Les données qui suivent montrent l'effet du temps sur le contenu d'hydrogène pour deux échantillons d'acier entreposés à 20 °C:

t	1	2	6	17	30
H	7,7	7,5	6,1	5,7	4,2
	8,4	8,1	6,8	5,3	4,5

où t désigne le temps (en heures) et H le contenu d'hydrogène (en parties par million). On propose le modèle

$$H = \beta_0 + \beta_1 \ln t + \varepsilon, \quad \text{où } \varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$$

(a) Estimer tous les paramètres du modèle.

(b) Calculer le tableau d'analyse de la variance et tester l'hypothèse nulle H_0: $\beta_1 = 0$ au seuil de signification $\alpha = 0{,}05$.

Question n$^{\text{o}}$ 5

On cherche à établir une relation entre le diamètre x (en centimètres) du filament d'une ampoule électrique et sa durée de vie Y (en heures). On dispose des données suivantes:

x	0,15	0,20	0,25	0,30	0,40	0,50	0,60	0,60
Y	120	165	204	238	296	373	410	403

x	0,60	0,70	0,70	0,80	0,80	0,80	0,90	1,00
Y	420	462	455	520	518	525	580	600

(a) Estimer les paramètres du modèle de régression linéaire simple.

(b) Calculer le tableau d'analyse de la variance et tester l'hypothèse nulle H_0: $\beta_1 = 0$ au seuil de signification $\alpha = 0{,}05$.

(c) Effectuer une analyse des résidus pour s'assurer que les hypothèses de base sont vérifiées.

Question n° 6

La loi des gaz parfaits reliant la pression P (en kilogrammes par centimètre carré) et le volume v (en centimètres cubes):

$$Pv^{\gamma} = c$$

peut s'écrire, après transformation, sous la forme

$$Y = \alpha - \gamma x$$

où $Y := \ln P$, $x := \ln v$ et $\alpha := \ln c$. Des observations particulières de v et P ont donné:

$$(v; P) \ = (50; 64{,}7)\,,\ (60; 51{,}3)\,,\ (70; 40{,}5)\,,\ (90; 25{,}9)\,,\ (100; 7{,}8)$$

Remarque. Calculs utiles:

$$\sum_{i=1}^{100} x_i \simeq 21{,}35984; \qquad \sum_{i=1}^{100} x_i^2 \simeq 91{,}57317; \qquad \sum_{i=1}^{100} x_i y_i \simeq 72{,}26249$$

$$\sum_{i=1}^{100} y_i \simeq 17{,}11712; \qquad \sum_{i=1}^{100} y_i^2 \simeq 61{,}40147$$

(a) Estimer les paramètres α et γ par la méthode des moindres carrés.

(b) Calculer le tableau d'analyse de la variance du modèle $Y = \alpha - \gamma x$ ainsi que le pourcentage de variation expliqué.

(c) Calculer un intervalle de confiance à 95 % pour la valeur moyenne de P lorsque $v = 80$.

Question n° 7

Le modèle de régression passant par l'origine s'écrit comme suit:

$$Y = \beta\, x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}(0, \sigma^2)$$

(a) Trouver l'estimateur du paramètre β par la méthode des moindres carrés, si l'on dispose de n observations du couple (x, Y).

(b) On suppose qu'une structure construite récemment s'enfonce dans le sol selon le modèle

$$Z = 10 - 10e^{-\beta x}$$

où x est l'âge (en mois) de la structure et Z est l'enfoncement (en centimètres). Trouver une transformation qui ramène le modèle ci-dessus au modèle de régression passant par l'origine.

Question n° 8

Dans le cadre d'une étude d'évaluation d'un processus de mesurage, un opérateur qui connaît bien l'appareil de mesure obtient deux mesures, X et Y, de 15 pièces d'un lot. Un processus de mesurage a une bonne *répétabilité* si le coefficient de corrélation, $r_{X,Y}$, de l'échantillon aléatoire du couple (X, Y) est supérieur à 0,80. Les données sont les suivantes:

X	34	56	6	50	33	43	49	17	54	24	35	46	10	51	25
Y	45	44	19	55	17	32	37	5	54	18	26	43	16	55	11

Remarque. Calculs utiles: $\bar{x} = 35{,}53$, $\bar{y} = 31{,}8$, $s_X \simeq 16{,}20$ et $s_Y \simeq 17{,}01$.

(a) Le processus de mesurage a-t-il une bonne répétabilité?

(b) Calculer la covariance $s_{X,Y} = \sum_{i=1}^{15} (x_i - \bar{x})\,(y_i - \bar{y})$ des observations particulières de X et Y.

(c) On considère les couples

$$(X_1, Y_1), \ldots, (X_{15}, Y_{15}), (Y_1, X_1), \ldots, (Y_{15}, X_{15})$$

Ceux-ci constituent 30 observations d'un couple de variables aléatoires que l'on notera (U, V).

(i) Expliquer pourquoi l'écart-type S_U des observations de U et celui des observations de V sont égaux.

(ii) On suppose que U et V présentent une distribution normale, de sorte que

$$E[V \mid U = u] = \beta_0 + \beta_1 u$$

Estimer, en utilisant les 30 observations particulières de (U, V), les paramètres β_0 et β_1 par la méthode de vraisemblance maximale.

Question n$^{\text{o}}$ 9

Durant les années 1840, le physicien écossais J. Forbes a mesuré en 17 endroits, dans les Alpes et en Écosse, la pression barométrique Y (en pouces de mercure) et la température x d'ébullition de l'eau (en degrés Fahrenheit). À cette époque, les baromètres étaient fragiles et il était beaucoup plus facile de déterminer la température d'ébullition de l'eau dans les régions montagneuses que d'employer un baromètre. À partir de cette température, on pouvait ensuite déterminer la pression barométrique. Les données de Forbes sont présentées ci-dessous:

x	194,5	194,3	197,0	198,4	199,4	199,9
Y	20,79	20,79	22,40	22,67	23,15	23,35

x	200,9	201,1	201,4	201,3	203,6	204,6
Y	23,89	23,99	24,02	24,01	25,14	26,57

x	209,5	208,6	210,7	211,9	212,2
Y	28,49	27,76	29,04	29,88	30,06

(a) Faire le graphique des points (x_i, y_i) et celui des points $(x_i, \ln y_i)$. Lequel des deux graphiques ressemble le plus à une droite?

(b) Effectuer une analyse de régression selon le modèle de régression linéaire simple. C'est-à-dire estimer les paramètres du modèle et faire le tableau d'analyse de la variance ainsi qu'une analyse des résidus.

(c) Effectuer une analyse de régression du modèle

$$\ln Y = \gamma_0 + \gamma_1 \, x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}(0, \sigma^2)$$

(d) Lequel des deux modèles est supérieur si l'on se base sur le pourcentage de variation expliqué par le modèle et le comportement des résidus?

Question n$^{\text{o}}$ 10

Soit $(x_1, y_1), \ldots, (x_{10}, y_{10})$ 10 observations particulières d'un couple de variables aléatoires (X, Y). On a: $\sum_{i=1}^{10} x_i = \sum_{i=1}^{10} y_i = 55, s_X = s_Y = 3{,}03$ et $\sum_{i=1}^{10} x_i y_i = 380$.

(a) Calculer le coefficient de corrélation de l'échantillon, $r_{X,Y}$.

(b) Calculer $\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2$.

Question n° 11

Une expérience semblable à celle de Forbes, décrite à la question n° 9, a été réalisée par J. Hooker dans l'Himalaya. Ses données sont les suivantes:

x	Y	x	Y	x	Y
210,8	29,211	193,6	21,212	185,6	17,062
210,2	28,559	193,4	20,480	184,6	16,881
208,4	27,972	191,4	19,758	184,1	16,959
202,5	24,697	191,1	19,490	184,1	16,817
200,6	23,726	190,6	19,386	183,2	16,385
200,1	23,369	189,5	18,869	182,4	16,235
199,5	23,030	188,8	18,356	181,9	16,106
197,0	21,892	188,5	18,507	181,9	15,928
196,4	21,928	186,0	17,221	181,0	15,919
196,3	21,654	185,7	17,267	180,6	15,376
195,6	21,605				

(a) Effectuer une analyse de régression du modèle

$$\ln Y = \gamma_0 + \gamma_1\, x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}(0, \sigma^2)$$

(b) Comparer les résultats avec ceux de la partie (c) de l'exercice n° 9.

(c) Effectuer une analyse de régression du modèle en (a) en utilisant les 48 données de Hooker et Forbes réunies.

Question n° 12

Dans un problème de régression linéaire simple, on a trouvé que le paramètre β_1 n'était pas significativement différent de zéro.

(a) On propose alors le modèle $Y = \beta_0 + \varepsilon$, où $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$. Estimer β_0 par la méthode des moindres carrés.

(b) On propose aussi le modèle $Y = \beta_0 + \beta_1 x^2 + \varepsilon$. Trouver la valeur $\hat{\beta}_0$ de β_0 qui minimise la fonction $SC\left(\beta_0, \hat{\beta}_1\right) := \sum_{i=1}^{n} \left(Y_i - \beta_0 - \hat{\beta}_1 x_i^2\right)^2$.

Question n° 13

Le voltage U (en volts) d'un condensateur chargé initialement à u_0 est, au bout de t secondes, donné par l'équation

$$U = u_0 e^{-\beta t}$$

Une expérience a donné les résultats suivants:

t	0	1	2	3	4	5	6	7	8	9	10
U	100	75	55	40	30	20	15	10	10	5	5

(a) Effectuer une transformation pour obtenir un modèle linéaire par rapport aux paramètres.

(b) Effectuer l'ajustement du nouveau modèle par la méthode des moindres carrés.

(c) Calculer des intervalles de confiance à 95 % pour u_0 et β.

Question n° 14

On propose le modèle $Y = \beta_0 + \beta_1 x + \varepsilon$, où $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$, reliant les variables x et Y. On a recueilli les informations suivantes: $\sum_{i=1}^{5} y_i = 10$, $\sum_{i=1}^{5} y_i^2 = 50$ et la somme des carrés SC_R égale 25.

(a) Calculer le coefficient de détermination, R^2.

(b) Calculer $\hat{\sigma}^2$.

Question n° 15

La loi de Boyle relie la pression P d'un gaz au volume v selon l'équation

$$Pv^{\alpha} = \beta$$

où α et β sont deux constantes. Effectuer l'ajustement du modèle à l'aide des données suivantes:

v	0,2	1,0	0,8	1,0	1,6	0,4
P	0,6	1,0	1,5	2,0	2,5	3,5

Question n° 16

On considère le modèle $Y = \beta_0 x^{\beta_0 \beta_1}$.

(a) Effectuer une transformation pour linéariser le modèle. Donner les nouvelles variables Y' et x', de même que les nouveaux paramètres β_0' et β_1', en fonction des anciens.

(b) Un tableau d'analyse de la variance a donné une statistique f_0 égale à 12. Peut-on alors rejeter, au seuil de signification $\alpha = 0{,}01$, l'hypothèse H_0: $\beta_1' = 0$, si l'on dispose de 10 observations particulières de (x, Y)? Donner la valeur du centile utilisé pour effectuer le test.

Question n^{o} 17 (voir la référence [2])

Un fabricant de vis désire renseigner ses clients au sujet de la relation entre la longueur nominale x et la longueur réelle Y de ses vis (en pouces). On a recueilli les valeurs suivantes, prises par des observations indépendantes de Y:

x	1/4	1/4	1/4	1/2	1/2	1/2
Y	0,262	0,262	0,245	0,496	0,512	0,490

x	3/4	3/4	3/4	1	1	1
Y	0,743	0,744	0,751	0,976	1,010	1,004

x	5/4	5/4	5/4	3/2	3/2	3/2
Y	1,265	1,254	1,252	1,498	1,518	1,504

x	7/4	7/4	7/4	2	2	2
Y	1,738	1,759	1,750	2,005	1,992	1,992

On propose le modèle de régression linéaire simple:

$$Y = \beta_0 + \beta_1\, x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}(0, \sigma^2)$$

(a) Estimer tous les paramètres du modèle.

(b) Obtenir un intervalle de confiance à 95 % pour la valeur moyenne de Y lorsque $x = 1$ en utilisant uniquement les trois observations particulières de Y pour cette valeur de x.

(c) Obtenir un intervalle de confiance à 95 % pour la valeur moyenne de Y lorsque $x = 1$ en utilisant toutes les données disponibles.

(d) Calculer le "coefficient de corrélation" des couples de données (x, y).

Questions à choix multiple

Question n^{o} 1

On étudie la relation entre le nombre x d'années écoulées depuis l'obtention de leur diplôme de premier cycle et le salaire annuel Y (en milliers de dollars) des ingénieurs. Les valeurs prises par les observations dans un échantillon aléatoire de taille $n = 10$ ont permis de constituer le tableau suivant:

x	1	2	3	4	5
Y	28	30	40	55	45
	31	35	36	40	60

On propose le modèle $Y = \beta_0 + \beta_1 x^2 + \varepsilon$, où $\varepsilon \sim \mathrm{N}\left(0, \sigma^2\right)$. On a alors: $SC_T = 1016$ et $\hat{\sigma}^2 \simeq 33{,}5$.

(A) Estimer le paramètre β_0 par la méthode des moindres carrés.

(a) 0 (b) 1 (c) 6,1 (d) 21,7 (e) 29

(B) Calculer le pourcentage de variation expliqué par le modèle proposé.

(a) 54,2 % (b) 73,2 % (c) 73,6 % (d) 85,6 % (e) 85,8 %

(C) Estimer le salaire d'un ingénieur qui possède 10 ans d'expérience, selon le modèle proposé.

(a) 100.000 (b) 104.000 (c) 105.000 (d) 121.700 (e) 129.000

Question n° 2

La vitesse réelle des voitures de trois marques, lorsque leurs indicateurs de vitesse marquent 100 km/h, a été mesurée pour quatre voitures de chaque marque:

Marque 1	105	103	108	101
Marque 2	101	97	99	104
Marque 3	102	96	98	103

On propose le modèle $V = \beta_0 + \beta_1 x + \epsilon$, où V est la vitesse réelle, x la vitesse donnée par l'indicateur de vitesse et $\epsilon \sim \mathrm{N}\left(0, \sigma^2\right)$. Estimer le paramètre β_1 par la méthode des moindres carrés.

(a) 0 (b) 1 (c) 1,1 (d) ∞ (e) indéterminé

Question n° 3

On s'intéresse à la relation entre deux variables aléatoires, X et Y. On a recueilli cinq couples d'observations particulières (x_i, y_i). Les résultats sont résumés ci-dessous:

$$\sum_{i=1}^{5} x_i = 5, \quad \sum_{i=1}^{5} y_i = 15, \quad \sum_{i=1}^{5} x_i^2 = 55, \quad \sum_{i=1}^{5} y_i^2 = 51, \quad \sum_{i=1}^{5} x_i y_i = 51$$

(A) Calculer le coefficient de corrélation empirique, $r_{X,Y}$.

(a) 0 (b) 0,194 (c) 0,4 (d) 0,775 (e) 0,8

(B) Supposons que $X \sim \mathrm{N}\left(\mu_X, \sigma_X^2\right)$ et $Y \sim \mathrm{N}\left(\mu_Y, \sigma_Y^2\right)$. Estimer la moyenne de Y étant donné que $X = 0$.

(a) -3 (b) 0 (c) 1,2 (d) 3 (e) 4,8

(C) Si, avec d'autres données, on a obtenu $R^2 = 0{,}04$ et $\hat{\beta}_1 > 0$, quelle est alors la valeur de la statistique utilisée pour tester l'hypothèse H_0: $\rho = 0$ contre H_1: $\rho \neq 0$?

(a) $-0{,}577$ (b) $-0{,}069$ (c) 0,069 (d) 0,354 (e) 0,577

Question n$^{\text{o}}$ 4

On considère les données suivantes:

x	1	1,5	2	2,5	3
Y	13	16	18	18	19

On propose le modèle de régression curviligne

$$Y = \beta_0 + \beta_1 \ln x + \epsilon, \quad \text{où } \epsilon \sim \mathrm{N}\big(0, \sigma^2\big)$$

On calcule

$$\sum_{i=1}^5 \ln x_i \simeq 3{,}1135; \qquad \sum_{i=1}^5 y_i = 84; \qquad \sum_{i=1}^5 y_i \ln x_i \simeq 56{,}3309$$

$$\sum_{i=1}^5 (\ln x_i)^2 \simeq 2{,}6914; \quad \sum_{i=1}^5 y_i^2 = 1434$$

(A) Estimer le paramètre β_1 par la méthode des moindres carrés.

(a) 2,80 (b) 5,35 (c) 11,20 (d) 13,47 (e) 17,12

(B) Calculer la valeur de la statistique F_0 utilisée pour tester l'hypothèse nulle H_0: $\beta_1 = 0$.

(a) 4,11 (b) 16,92 (c) 50,77 (d) 72,30 (e) 75,20

(C) Quel est le pourcentage de variation expliqué par le modèle?

(a) 5,58 % (b) 12,62 % (c) 80,0 % (d) 88,22 % (e) 94,42 %

8

Fiabilité

Dans plusieurs domaines appliqués, en particulier dans la plupart des disciplines du génie, il est important d'être en mesure de calculer la probabilité qu'un certain dispositif ou système soit en état de fonctionnement à un instant donné, ou pendant un temps déterminé. Nous avons déjà considéré plusieurs exercices sur la *fiabilité* aux chapitres 2 à 4. Au chapitre 2, il était sous-entendu que l'on calculait la fiabilité d'un système à un instant donné t_0, en supposant que la fiabilité de chacun de ses composants à cet instant était connue. Pour pouvoir calculer la probabilité qu'une machine fonctionne sans panne pendant une certaine quantité de temps, il est nécessaire de connaître la distribution de sa durée de vie ou de la durée de vie de ses composants. Cela devient un problème sur les variables ou les vecteurs aléatoires. Dans ce chapitre, nous présentons en détail les principaux concepts de la fiabilité.

8.1 Notions de base

Il y a plusieurs interprétations possibles du mot *fiabilité*. Dans ce livre, il correspond toujours à la probabilité qu'un système fonctionne correctement à un instant donné ou pendant une période donnée. De plus, dans ce chapitre, nous nous intéressons surtout à la fiabilité pendant un certain intervalle de temps $[0, t]$.

Définition 8.1.1. *Soit X une variable aléatoire non négative représentant la durée de vie (ou le temps jusqu'à une panne) d'un système ou d'un dispositif. La probabilité*

$$R(x) = P[X > x] \quad [= 1 - F_X(x)] \quad \textit{pour } x \geq 0$$

est appelée **fonction de fiabilité** *ou* **fonction de survie** *du système.*

Remarques.
(i) La fonction $R(x)$ peut aussi être notée $S(x)$. La notation $\bar{F}_X(x) = 1 - F_X(x)$ est également utilisée.

(ii) La plupart du temps, on suppose que la variable aléatoire X est continue. Cependant, dans certaines applications, la durée de vie est mesurée en nombre de cycles. Par conséquent, X est alors une variable aléatoire discrète (prenant des valeurs entières non négatives, plus précisément). De plus, si l'on accepte la possibilité que le dispositif puisse être défectueux, alors X pourrait prendre la valeur 0 et être une variable aléatoire de *type mixte*.

(iii) Toutes les distributions discrètes considérées à la section 3.2 pourraient servir de modèles en fiabilité. Dans le cas des distributions continues, nous devons nous limiter à celles qui sont toujours non négatives. Par conséquent, la distribution normale ne peut pas être un modèle *exact* en fiabilité. Malgré tout, selon la valeur des paramètres μ et σ, elle peut être un bon modèle *approximatif* pour la durée de vie d'une machine. De plus, on peut considérer la distribution normale *tronquée*, définie pour $x \geq 0$.

Une mesure utile de la fiabilité d'un système est sa *durée de vie moyenne* $E[X]$, soit le temps moyen de fonctionnement jusqu'à une panne du système.

Définition 8.1.2. *Le symbole $MTTF$ (***Mean Time To Failure***, en anglais) représente l'espérance mathématique de la durée de vie X d'un système. Si le système peut être réparé, on définit aussi les symboles $MTBF$ pour* temps moyen entre deux pannes *(***Mean Time Between Failures***), et $MTTR$ pour* temps moyen de réparation *(***Mean Time To Repair***). On a: $MTBF = MTTF + MTTR$.*

Remarques.
(i) Supposons qu'on s'intéresse à la durée de vie X d'une voiture. Il est évident que, sauf dans le cas d'une panne réellement majeure, la voiture sera réparée quand elle tombera en panne. Lorsqu'on calcule la quantité $MTBF$, on suppose que, après avoir été *réparé*, un système est *aussi bon que neuf*. Bien sûr, dans le cas d'une voiture, cela n'est pas exact, car les voitures *vieillissent* et s'usent.

(ii) Pour différencier les pannes *critiques* des non critiques, on peut utiliser le terme plus précis *Mean Time Between Critical Failures* ($MTBCF$), pour *temps moyen entre deux pannes critiques*. Alors le symbole $MTBF$ pourrait être interprété comme le temps moyen entre deux pannes quelconques, c'est-à-dire critiques ou non. Dans le contexte d'un système informatique de transmission

de données, on a aussi le *Mean Time Between System Aborts* ($MTBSA$), soit le *temps moyen entre deux arrêts prématurés.*

Pour calculer la durée de vie moyenne d'un système, on peut bien sûr utiliser la définition de l'espérance mathématique d'une variable aléatoire. Cependant, il est parfois plus simple de procéder comme dans la proposition suivante.

Proposition 8.1.1. *Soit X une variable aléatoire non négative. Alors on a:*

$$E[X] = \begin{cases} \sum_{k=0}^{\infty} P[X > k] = \sum_{k=0}^{\infty} R(k) \ \textit{si } X \in \{0, 1, \ldots\} \\ \int_0^{\infty} P[X > x]dx = \int_0^{\infty} R(x)dx \ \textit{si } X \in [0, \infty) \end{cases}$$

Preuve. Considérons d'abord le cas où X est une variable aléatoire qui prend des valeurs entières non négatives. On a:

$$\begin{aligned} E[X] &:= \sum_{j=0}^{\infty} jP[X = j] = \sum_{j=1}^{\infty} jP[X = j] = \sum_{j=1}^{\infty} \sum_{k=1}^{j} P[X = j] \\ &= \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} P[X = j] = \sum_{k=1}^{\infty} P[X \geq k] = \sum_{k=0}^{\infty} P[X > k] \end{aligned}$$

De façon similaire, si X (est continue et) prend ses valeurs dans l'intervalle $[0, \infty)$, on peut écrire que

$$\begin{aligned} E[X] &:= \int_0^{\infty} t f_X(t)dt = \int_0^{\infty} \int_0^{t} f_X(t)dxdt \\ &= \int_0^{\infty} \int_x^{\infty} f_X(t)dtdx = \int_0^{\infty} P[X > x]dx \end{aligned}$$

■

Remarques.
(i) Il n'est pas nécessaire que la variable aléatoire discrète X prenne toutes les valeurs entières non négatives. Il suffit que l'ensemble des valeurs possibles de X soit inclus dans $\{0, 1, \ldots\}$. De même, dans le cas continu, X doit prendre ses valeurs dans l'intervalle $[0, \infty)$.

(ii) Souvent, les formules dans la proposition ne simplifient pas le calcul de l'espérance mathématique de X. Par exemple, il est plus compliqué de calculer

la moyenne d'une variable aléatoire qui présente une distribution de Poisson à partir de la première formule qu'à partir de la définition.

Exemple 8.1.1. Soit X une variable aléatoire qui présente une distribution géométrique de paramètre p appartenant à l'intervalle $(0, 1)$. Ses valeurs possibles sont les entiers $1, 2, \ldots$, et sa moyenne est égale à $1/p$. Nous avons vu (à la page 81) que

$$P[X > k] = (1-p)^k \quad \text{pour } k = 0, 1, \ldots$$

Il s'ensuit qu'on a bien:

$$E[X] = \sum_{k=0}^{\infty} (1-p)^k = \frac{1}{1-(1-p)} = \frac{1}{p}$$

♢

Exemple 8.1.2. Si $X \sim \text{Exp}(\lambda)$, on trouve (voir l'exemple 3.5.2) que

$$P[X > x] = \int_x^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda x} \quad \text{pour } x \geq 0$$

De là,

$$E[X] = \int_0^{\infty} e^{-\lambda x} dx = -\frac{e^{-\lambda x}}{\lambda}\Big|_0^{\infty} = \frac{1}{\lambda}$$

♢

Au chapitre 4, nous avons défini diverses fonctions conditionnelles, par exemple la fonction $f_X(x \mid Y = y)$, où (X, Y) est un vecteur aléatoire continu. Nous avons aussi défini des fonctions du type $f_X(x \mid A_X)$, où A_X est un événement qui n'implique que la variable aléatoire X. Une certaine *fonction de densité conditionnelle* est importante en fiabilité.

Définition 8.1.3. *Supposons que la durée de vie T d'un système est une variable aléatoire continue et non négative. Le* **taux de panne** *(ou* **taux de défaillance***) $r(t)$ du système est défini par*

$$r(t) = f_T(t \mid T > t) := \lim_{s \downarrow t} \frac{d}{ds} F_T(s \mid T > t) \quad \textit{pour } t \geq 0$$

Remarques.
(i) La fonction $r(t)$, multipliée par dt, peut être interprétée comme étant la probabilité qu'une machine, qui est âgée de t unités de temps et qui fonctionne encore, tombe en panne dans l'intervalle $(t, t+dt]$. En effet, on a:

$$f_T(t \mid T > t) = \lim_{dt \downarrow 0} \frac{P[t < T \leq t + dt \mid T > t]}{dt}$$

(ii) On suppose que la *fonction de répartition conditionnelle* $F_T(s \mid T > t)$ est dérivable en $s \in (t, t+dt]$.
(iii) On doit prendre la limite lorsque s décroît vers t dans la définition, parce que $F_T(t \mid T > t) \equiv P[T \leq t \mid T > t) = 0$. Cependant, on n'a pas à prendre une limite pour calculer $f_T(s \mid T > t)$ à partir de $F_T(s \mid T > t)$, pour $s > t$.

Proposition 8.1.2. *On a:*

$$r(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{R'(t)}{R(t)} \quad \textit{pour } t \geq 0$$

Preuve. Par définition,

$$F_T(s \mid T > t) = P[T \leq s \mid T > t] = \frac{P[\{T \leq s\} \cap \{T > t\}]}{P[T > t]}$$
$$= \begin{cases} 0 & \text{si } s \leq t \\ \dfrac{F_T(s) - F_T(t)}{1 - F_T(t)} & \text{si } s > t \end{cases}$$

De là,

$$f_T(s \mid T > t) := \frac{d}{ds} F_T(s \mid T > t) = \frac{f_T(s)}{1 - F_T(t)} \quad \text{si } s > t$$

En prenant la limite lorsque s décroît vers t, on obtient que

$$r(t) := f_T(t \mid T > t) = \frac{f_T(t)}{1 - F_T(t)}$$

Finalement, puisque

$$R'(t) = \frac{d}{dt}[1 - F_T(t)] = -f_T(t)$$

on a aussi:

$$r(t) = -\frac{R'(t)}{R(t)} \quad \text{pour } t \geq 0$$

■

Remarque. Dans le cas discret, le taux de panne est donné par

$$r(k) = \frac{p_X(k)}{\sum_{j=k}^{\infty} p_X(j)} \quad \text{pour } k = 0, 1, \ldots$$

Notons que $0 \leq r(k) \leq 1$ pour tout k, tandis que $r(t) \geq 0$ dans le cas continu.

Exemple 8.1.3. Un des modèles les plus utilisés en fiabilité est la distribution exponentielle, principalement en raison de sa propriété de non-vieillissement (voir la page 94). Cette propriété implique que pour un système dont la durée de vie possède une distribution exponentielle, le taux de panne est *constant*. En effet, si $X \sim \text{Exp}(\lambda)$, on a (voir l'exemple précédent):

$$r(t) = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \quad \text{pour } t \geq 0$$

En pratique, cela n'est généralement pas réaliste. Il y a cependant des applications pour lesquelles cela est acceptable. Par exemple, il semble que la distribution de la durée de vie d'un fusible électrique qui ne peut pas fondre que partiellement soit approximativement exponentielle. Le temps entre les pannes d'un système constitué d'un très grand nombre de composants indépendants reliés en série peut aussi présenter approximativement une distribution exponentielle, si l'on suppose, en particulier, que chaque fois qu'un composant tombe en panne il est immédiatement remplacé par un composant neuf. Toutefois, dans la plupart des cas, la distribution exponentielle ne devrait être utilisée que pour t dans un intervalle fini $[t_1, t_2]$. ♢

Exemple 8.1.4. La distribution géométrique est l'équivalent de la distribution exponentielle en temps discret. Elle possède aussi la propriété de non-vieillissement. Puisque $P[X \geq k] = P[X > k-1]$, on calcule (voir l'exemple 8.1.1)

$$r(k) = \frac{(1-p)^{k-1}p}{(1-p)^{k-1}} = p \quad \text{pour } k = 1, 2, \ldots$$

Ainsi, le taux de panne $r(k)$ est une constante dans ce cas également, comme on pouvait s'y attendre. ♢

Le taux de panne d'une distribution donnée est un bon indicateur de la valeur de cette distribution comme modèle en fiabilité. Dans la plupart des applications, $r(t)$ devrait être une fonction strictement croissante de t, au moins lorsque t est assez grand.

Définition 8.1.4. *Si la variable aléatoire X est telle que son taux de panne $r_X(t)$ ou $r_X(k)$ est une fonction* croissante *(respectivement* décroissante*) de t ou k, alors on dit que X possède une distribution à* **taux de panne croissant** *(respectivement* **taux de panne décroissant***).*

Notation. Nous utiliserons les symboles *IFR* pour *taux de panne croissant* (*Increasing Failure Rate*, en anglais) et *DFR* pour *taux de panne décroissant* (*Decreasing Failure Rate*, en anglais).

Maintenant, en utilisant la proposition 8.1.2, on obtient que

$$\int_0^t r(s)ds = -\int_0^t \frac{R'(s)}{R(s)}ds = -\ln R(t) + \ln R(0)$$

De plus, la variable aléatoire T (ayant un taux de panne $r(t)$) étant continue et non négative, on peut écrire que $R(0) := P[T > 0] = 1$. De là, on peut énoncer la proposition suivante.

Proposition 8.1.3. *Il y a une correspondance biunivoque entre les fonctions $R(t)$ et $r(t)$:*

$$R(t) = \exp\left\{ -\int_0^t r(s)ds \right\}$$

Remarque. La proposition implique que la distribution exponentielle est la *seule* distribution continue ayant un taux de panne constant.

Exemple 8.1.5. On peut montrer que le taux de panne $r(t)$ d'une distribution lognormale commence à zéro (parce que $\lim_{t\downarrow 0} f_T(t) = 0$), puis qu'il augmente jusqu'à un maximum, et ensuite

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

Par conséquent, on doit conclure que la distribution lognormale *n'est pas* un bon modèle pour la durée de vie d'un dispositif qui est sujet à l'usure, du moins pas pour t grand. En effet, le taux de panne devrait généralement augmenter avec t, comme mentionné ci-dessus. ♢

Exemple 8.1.6. La distribution normale $\mathrm{N}(\mu, \sigma^2)$ ne devrait pas être utilisée pour modéliser la durée de vie d'un système, à moins que les paramètres μ et σ soient tels que la probabilité que la variable aléatoire prenne une valeur négative soit négligeable. Pour n'importe quelles valeurs de μ et σ, on peut définir la distribution normale *tronquée* comme suit:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma\, c} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad \text{pour } x \geq 0$$

où c est une constante telle que $\int_0^\infty f_X(x)dx = 1$. C'est-à-dire que

$$c = \left[\int_0^\infty f_Y(y)dy\right]^{-1} = \frac{1}{1 - \Phi(-\mu/\sigma)}$$

où $Y \sim \mathrm{N}(\mu, \sigma^2)$. On peut écrire que $X \equiv Y \mid \{Y \geq 0\}$. Notons que si $\mu = 0$, alors $c = 2$.

On trouve que le taux de panne d'une distribution normale tronquée est strictement croissant, ce qui en fait un modèle intéressant pour plusieurs applications. ♢

Exemple 8.1.7. La distribution de Weibull (voir la page 95) est un modèle réellement important en fiabilité et en analyse de la *fatigue*. On a:

$$R(t) = \int_t^\infty \lambda\beta x^{\beta-1} \exp\left(-\lambda x^\beta\right) dx = \exp\left(-\lambda t^\beta\right)$$

Il s'ensuit que

$$r(t) = \frac{\lambda\beta t^{\beta-1} \exp\left(-\lambda t^\beta\right)}{\exp\left(-\lambda t^\beta\right)} = \lambda\beta t^{\beta-1} \quad \text{pour } t \geq 0$$

Par conséquent, la distribution de Weibull est DFR si $\beta < 1$ et IFR si $\beta > 1$. Lorsque $\beta = 1$, on retrouve la distribution exponentielle. ♢

Bien qu'il soit vrai que, pour des valeurs de t assez grandes, le taux de panne $r(t)$ devrait augmenter lorsque t augmente, dans plusieurs situations, ce taux est d'abord une fonction décroissante de t. Par exemple, le *taux de mortalité* des enfants diminue effectivement au départ. En effet, il y a un plus grand risque qu'un bébé meure à la naissance ou peu après que lorsqu'il est âgé de six mois, en particulier. Quand l'enfant vieillit, le taux de mortalité est plus ou moins constant pendant un certain temps, tandis qu'il croît chez les adultes. Par

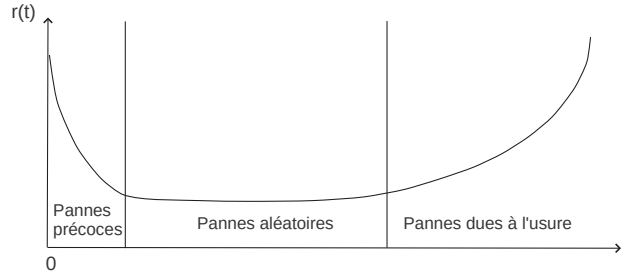

Fig. 8.1. Taux de panne ayant la forme d'une baignoire

conséquent, la fonction $r(t)$ ressemble à une *baignoire* (voir la figure 8.1). Comme nous l'avons mentionné ci-dessus, la distribution exponentielle ne devrait être utilisée que pour t tel que $t_1 \leq t \leq t_2 < \infty$. Elle est valable pour la portion relativement plate de la baignoire.

Supposons que la durée de vie X est définie comme suit:

$$X = c_1 X_1 + c_2 X_2 + c_3 X_3$$

où X_i présente une distribution de Weibull, et les constantes c_i sont positives, pour $i = 1, 2, 3$, et telles que $c_1 + c_2 + c_3 = 1$. La combinaison linéaire de variables aléatoires de Weibull est alors une *distribution de Weibull mixte*. La forme de la baignoire peut être obtenue en choisissant le paramètre β de X_1 inférieur à 1, celui de X_2 égal à 1, et celui de X_3 supérieur à 1. Notons que X_2 est en fait une variable aléatoire exponentielle.

On est parfois intéressé à connaître la probabilité qu'un système tombe en panne pendant un intervalle de temps particulier.

Définition 8.1.5. *Le* **taux de panne** *d'un système* **dans un intervalle** $(t_1, t_2]$ *est noté* $FR(t_1, t_2)$ *et est défini par*

$$FR(t_1, t_2) = \frac{P[t_1 < T \leq t_2 \mid T > t_1]}{t_2 - t_1} = \frac{R(t_1) - R(t_2)}{R(t_1)} \frac{1}{(t_2 - t_1)}$$

pour $0 \leq t_1 < t_2 < \infty$*, où* T *est une variable aléatoire continue.*

Remarque. On a:

$$r(t_1) = \lim_{t_2 \downarrow t_1} FR(t_1, t_2)$$

Exemple 8.1.8. Supposons que $T \sim \text{Exp}(\lambda)$. D'abord, on calcule la fonction de densité conditionnelle $f_T(t \mid T > t_1)$. On trouve que

$$f_T(t \mid T > t_1) = \frac{f_T(t)}{P[T > t_1]} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t_1}} = \lambda e^{-\lambda(t-t_1)} \quad \text{pour } t > t_1$$

(et $f_T(t \mid T > t_1) = 0$ pour $t \leq t_1$). Il s'ensuit que

$$P[t_1 < T \leq t_2 \mid T > t_1] = \int_{t_1}^{t_2} \lambda e^{-\lambda(t-t_1)} dt = -e^{-\lambda(t-t_1)} \Big|_{t_1}^{t_2} = 1 - e^{-\lambda(t_2-t_1)}$$

De là, on a:

$$FR(t_1, t_2) = \frac{1 - e^{-\lambda(t_2-t_1)}}{t_2 - t_1}$$

En fait, on aurait pu obtenir ce résultat immédiatement à partir de la fonction de fiabilité $R(t) = e^{-\lambda t}$. Cependant, le but était de donner la formule de la fonction de densité de la distribution exponentielle *décalée*. Une distribution décalée peut être utilisée comme modèle en fiabilité dans la situation suivante: supposons qu'une personne achète un appareil pour lequel il y a une période de garantie de $t_1 > 0$ unité(s) de temps. Alors cette personne est certaine que l'appareil en question durera au moins t_1 unité(s) de temps (l'appareil peut être réparé ou remplacé s'il tombe en panne avant la fin de la période de garantie).

Notons que la fonction $FR(t_1, t_2)$ ne dépend en fait que de la différence $t_2 - t_1$. Par conséquent, la probabilité que le système tombe en panne dans un intervalle donné ne dépend que de la longueur de cet intervalle, ce qui est une conséquence de la propriété de non-vieillissement de la distribution exponentielle.

Finalement, en utilisant la règle de l'Hospital, on obtient que

$$\lim_{t_2 \downarrow t_1} FR(t_1, t_2) = \lim_{t_2 \downarrow t_1} \frac{1 - e^{-\lambda(t_2-t_1)}}{t_2 - t_1} = \lim_{\epsilon \downarrow 0} \frac{1 - e^{-\lambda\epsilon}}{\epsilon} = \lim_{\epsilon \downarrow 0} \frac{e^{-\lambda\epsilon}\lambda}{1} = \lambda$$

comme cela devait être le cas. ♢

Exemple 8.1.9. Soit

$$f_T(t) = te^{-t} \quad \text{pour } t > 0$$

La variable aléatoire T présente une distribution gamma de paramètres $\alpha = 2$ et $\lambda = 1$. On calcule

$$R(t) = \int_t^\infty se^{-s} ds = -se^{-s} \Big|_t^\infty + \int_t^\infty e^{-s} ds = (t+1)e^{-t} \quad \text{pour } t > 0$$

où on a utilisé la règle de l'Hospital pour évaluer la limite $\lim_{s\to\infty} se^{-s}$. On a:

$$r(t) = \frac{te^{-t}}{(t+1)e^{-t}} = \frac{t}{t+1} = 1 - \frac{1}{t+1}$$

ce qui est une fonction croissante pour toute valeur de t. En fait, la distribution gamma est IFR pour n'importe quel $\alpha > 1$ (et DFR si $0 < \alpha < 1$).

À partir de la fonction $R(t)$, on déduit que

$$FR(t_1, t_2) = \left[1 - \frac{(t_2+1)}{(t_1+1)} e^{-(t_2-t_1)}\right] \frac{1}{t_2 - t_1}$$

♢

Pour terminer cette section, nous définissons une autre quantité d'intérêt en fiabilité.

Définition 8.1.6. *Le* **taux moyen de panne** *d'un système* **dans un intervalle** $[t_1, t_2]$ *est donné par*

$$AFR(t_1, t_2) = \frac{\int_{t_1}^{t_2} r(t)dt}{t_2 - t_1} = \frac{\ln[R(t_1)] - \ln[R(t_2)]}{t_2 - t_1}$$

Remarque. Cette quantité est un exemple de *moyenne temporelle.*

Exemple 8.1.10. Si T présente une distribution de Weibull, on a (voir l'exemple 8.1.7):

$$\int_{t_1}^{t_2} r(t)dt = \int_{t_1}^{t_2} \lambda\beta t^{\beta-1} dt = \lambda(t_2^\beta - t_1^\beta)$$

de sorte que

$$AFR(t_1, t_2) = \frac{\lambda(t_2^\beta - t_1^\beta)}{t_2 - t_1}$$

Si $\beta = 1$, on a que $AFR(t_1, t_2) \equiv \lambda$, tandis que $\beta = 2$ implique que $AFR(t_1, t_2)$ est égal à $\lambda(t_1 + t_2)$. Notons que, lorsque $\beta = 2$, le taux moyen de panne est trois fois supérieur dans l'intervalle $(t, 2t)$ que de 0 à t. ♢

8.2 Fiabilité des systèmes

Dans cette section, nous considérons des systèmes constitués d'au moins deux sous-systèmes ou composants qui peuvent être placés *en série* ou *en parallèle.*

Lorsque les composants sont placés en parallèle, il faut distinguer entre *redondance active* et *redondance passive* (ou *en attente*). On suppose que les composants qui constituent les systèmes considérés ne peuvent pas être réparés. Lorsqu'un système tombe en panne, il le demeure indéfiniment. En fait, il recommencerait à fonctionner si les composants en panne étaient remplacés par des composants neufs. Cependant, ici, nous nous intéressons seulement au temps écoulé jusqu'à la *première* panne du système.

8.2.1 Systèmes en série

Considérons n sous-systèmes qui fonctionnent indépendamment les uns des autres. Soit $R_k(t)$ la fonction de fiabilité du sous-système k, pour $k = 1, \ldots, n$. Si les sous-systèmes sont placés en série, chaque sous-système doit être opérationnel pour que le système fonctionne. Par conséquent, la durée de vie T du système est telle que

$$T > t \quad \Longleftrightarrow \quad T_k > t \quad \text{pour tout } k$$

où T_k est la durée de vie du sous-système k. Il s'ensuit que la fonction de fiabilité du système est donnée par (voir la proposition 8.1.3)

$$R(t) = \prod_{k=1}^{n} R_k(t) = \exp\left\{ -\int_0^t [r_1(s) + \cdots + r_n(s)]\, ds \right\}$$

Remarques.
(i) Au chapitre 2, nous aurions pu demander de calculer la probabilité qu'un système constitué de n composants indépendants connectés en série fonctionne à un instant donné t_0, étant donné que l'on connaît la fiabilité de chaque composant à t_0. Soit les événements S: le système fonctionne à l'instant t_0, et B_k: le composant k fonctionne à l'instant t_0. On a:

$$P[S] = P\left[\bigcap_{k=1}^{n} B_k\right] \overset{\text{ind.}}{=} \prod_{k=1}^{n} P[B_k]$$

On a posé ci-dessus l'hypothèse selon laquelle les composants ne peuvent pas être réparés, de sorte que cette formule est en fait un cas particulier de la formule précédente. En effet, on peut écrire que $P[S] = R(t_0)$ et $P[B_k] = R_k(t_0)$. On doit aussi supposer qu'aucun composant n'a été remplacé dans l'intervalle $(0, t_0)$.

(ii) Lorsque les composants qui constituent un système sont placés en série, on doit supposer qu'ils fonctionnent indépendamment les uns des autres, parce que

le système tombe en panne dès qu'un de ses composants cesse de fonctionner. Par exemple, supposons qu'il n'y a que deux composants, notés A et B. On ne peut pas poser qu'il y a une certaine probabilité $p_1(t)$ (respectivement $p_2(t)$) que le composant B soit actif à l'instant t si A est actif (respectivement en panne) à l'instant t. Si le composant A est en panne à l'instant t, alors le système a cessé de fonctionner lorsque A est tombé en panne et le système est demeuré en panne à partir de ce moment. De plus, en temps continu, la probabilité que les composants A et B tombent en panne exactement au même instant est nulle. De là, si l'on suppose qu'un composant ne peut pas tomber en panne lorsque le système ne fonctionne plus, si A est en panne, alors B doit être opérationnel. À moins que la durée de vie du composant B présente une distribution exponentielle (ou une distribution géométrique, en temps discret), lorsque le composant A est remplacé par un composant neuf, le composant B devait être remplacé également, si l'on désire que le système soit aussi bon que neuf. De façon similaire, on ne peut pas supposer que la durée de vie du composant A présente une distribution qui dépend de la durée de vie de B.

(iii) On peut écrire que

$$T = \min\{T_1, T_2, \ldots, T_n\}$$

Le *minimum* d'un ensemble de variables aléatoires est un cas spécial de ce qu'on appelle *statistiques d'ordre*.

Proposition 8.2.1. *Supposons que T_1 et T_2 sont des variables aléatoires qui présentent des distributions exponentielles indépendantes, de paramètres λ_1 et λ_2, respectivement. On a:*

$$T := \min\{T_1, T_2\} \sim \mathit{Exp}(\lambda_1 + \lambda_2)$$

Preuve. On calcule

$$P[T > t] \overset{\text{ind.}}{=} P[T_1 > t]P[T_2 > t] = e^{-\lambda_1 t}e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t} \quad \text{pour } t \geq 0$$

Il s'ensuit que

$$f_T(t) = \frac{d}{dt}\{1 - P[T > t]\} = \frac{d}{dt}\{1 - e^{-(\lambda_1+\lambda_2)t}\} = (\lambda_1 + \lambda_2)e^{-(\lambda_1+\lambda_2)t}$$

pour $t \geq 0$. ■

Remarque. La proposition peut être généralisée comme suit: si $T_k \sim \text{Exp}(\lambda_k)$, pour $k = 1, \ldots, n$, et si les T_k sont des variables aléatoires indépendantes, alors

$T := \min\{T_1, T_2, \ldots, T_n\} \sim \text{Exp}(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$. Par conséquent, si n composants indépendants dont les durées de vie présentent des distributions exponentielles sont placés en série, c'est comme s'il y avait un seul composant dont la durée de vie présente une distribution exponentielle de paramètre λ égal à la somme des λ_k. Notons que puisque $r_k(t) \equiv \lambda_k$, pour $k = 1, \ldots, n$, on a:

$$R(t) = \exp\left\{ -\int_0^t (\lambda_1 + \cdots + \lambda_n)\, ds \right\} = \exp[-(\lambda_1 + \cdots + \lambda_n)t] = e^{-\lambda t}$$

où $\lambda := \lambda_1 + \cdots + \lambda_n$.

Exemple 8.2.1. Si T_k présente une distribution uniforme sur l'intervalle $[0, 1]$, pour $k = 1, \ldots, n$, alors

$$R_k(t) = \int_t^1 1 ds = 1 - t \quad \text{pour } 0 \leq t \leq 1$$

Il s'ensuit que

$$R(t) = \prod_{k=1}^{n} R_k(t) = \prod_{k=1}^{n} (1 - t) = (1 - t)^n \quad \text{pour } 0 \leq t \leq 1$$

Puisque $T_k \leq 1$ pour tout k, on peut écrire que $R(t) = 0$ si $t > 1$. ♢

8.2.2 Systèmes en parallèle

Redondance active

Nous considérons maintenant des systèmes constitués d'au moins deux sous-systèmes connectés *en parallèle*. Supposons d'abord que tous les sous-systèmes, qui peuvent comprendre un ou plusieurs composants, fonctionnent à partir de l'instant initial $t = 0$. C'est ce que l'on appelle la *redondance active*. Le système en entier fonctionnera tant qu'il restera au moins un sous-système opérationnel. On peut écrire que la durée de vie T du système est le *maximum* des variables aléatoires $T_1, \ldots, T_n$, où n est le nombre de sous-systèmes placés en parallèle. Il s'ensuit que

$$T \leq t \quad \Longleftrightarrow \quad T_k \leq t \quad \text{pour } k = 1, \ldots, n$$

De là, si les sous-systèmes fonctionnent indépendamment les uns des autres, on a:

$$R(t) = 1 - \prod_{k=1}^{n} [1 - R_k(t)]$$

Remarque. Lorsque les sous-systèmes sont placés en parallèle, on peut considérer le cas où ils ne fonctionnent pas indépendamment. En fait, ce cas a déjà été considéré au chapitre 2, en particulier dans l'exemple 2.4.1.

Exemple 8.2.2. Un dispositif comprend deux composants placés en parallèle et fonctionnant indépendamment l'un de l'autre. La durée de vie T_k du composant k présente une distribution exponentielle de paramètre λ_k, pour $k = 1, 2$. Il s'ensuit que

$$\begin{aligned} P[T \le t] &= P[\{T_1 \le t\} \cap \{T_2 \le t\}] \overset{\text{ind.}}{=} P[T_1 \le t] P[T_2 \le t] \\ &= (1 - e^{-\lambda_1 t})(1 - e^{-\lambda_2 t}) \quad \text{pour } t \ge 0 \end{aligned}$$

où T est la durée de vie totale du système. De là, la fonction de fiabilité du système est

$$R(t) = e^{-\lambda_1 t} + e^{-\lambda_2 t} - e^{-(\lambda_1 + \lambda_2)t}$$

Notons que le maximum de deux variables aléatoires exponentielles indépendantes ne présente pas une distribution exponentielle (même pas dans le cas où $\lambda_1 = \lambda_2$), car

$$f_T(t) = \frac{d}{dt} P[T \le t] = \lambda_1 e^{-\lambda_1 t} + \lambda_2 e^{-\lambda_2 t} - (\lambda_1 + \lambda_2) e^{-(\lambda_1 + \lambda_2)t} \quad \text{pour } t \ge 0$$

On a:

$$E[T] = \int_0^\infty R(t)\,dt = \frac{1}{\lambda_1} + \frac{1}{\lambda_2} - \frac{1}{\lambda_1 + \lambda_2}$$

Dans le cas particulier où $\lambda_1 = \lambda_2 = \lambda$, on obtient que $E[T] = 1{,}5/\lambda$. C'est-à-dire que le fait d'installer deux composants identiques en parallèle, dans cet exemple, augmente de 50 % le temps moyen de fonctionnement jusqu'à une panne du dispositif. ♢

Exemple 8.2.3. Supposons que, dans l'exemple précédent, la probabilité que le composant n° 2 soit actif à un instant fixé $t_0 > 0$ est égale à $e^{-\lambda_{21} t_0}$ si le composant n° 1 est aussi actif à l'instant t_0, et à $e^{-\lambda_{22} t_0}$ si le composant n° 1 est en panne à l'instant t_0. On peut alors écrire (parce que la distribution exponentielle est continue, de sorte que $P[T_1 = t_0] = 0$) que

$$P[T \le t_0] = P[\{T_1 \le t_0\} \cap \{T_2 \le t_0\}] = P[T_2 \le t_0 \mid T_1 \le t_0]P[T_1 \le t_0]$$
$$= (1 - e^{-\lambda_{22} t_0})(1 - e^{-\lambda_1 t_0})$$

De plus, on a:

$$P[T_2 \le t_0] = P[T_2 \le t_0 \mid T_1 \le t_0]P[T_1 \le t_0] + P[T_2 \le t_0 \mid T_1 > t_0]P[T_1 > t_0]$$
$$= (1 - e^{-\lambda_{22} t_0})(1 - e^{-\lambda_1 t_0}) + (1 - e^{-\lambda_{21} t_0})e^{-\lambda_1 t_0}$$

En général, la constante λ_{22} devrait en fait être une fonction de l'instant exact auquel le composant n° 1 est tombé en panne. Dans l'exemple 2.4.1, nous avons fourni les valeurs numériques des probabilités que le composant B fonctionne à un instant non spécifié, étant donné que le composant A fonctionne, ou ne fonctionne pas, à cet instant. Notons que la somme $e^{-\lambda_{21} t_0} + e^{-\lambda_{22} t_0}$ peut prendre n'importe quelle valeur dans l'intervalle $[0, 2]$.

Remarque. En *conditionnant* sur l'instant où le composant n° 1 est tombé en panne, on peut montrer que

$$P[T_2 > t] = \int_0^\infty P[T_2 > t \mid T_1 = \tau] f_{T_1}(\tau) d\tau$$
$$= \int_0^t P[T_2 > t \mid T_1 = \tau] f_{T_1}(\tau) d\tau + \int_t^\infty P[T_2 > t \mid T_1 = \tau] f_{T_1}(\tau) d\tau$$

Supposons que T_2 présente une distribution exponentielle de paramètre λ_2 tant que le composant n° 1 fonctionne, et une distribution exponentielle de paramètre λ_3 (qui devrait être supérieur à λ_2) à partir du moment où le composant n° 1 tombe en panne. Alors, par la propriété de non-vieillissement de la distribution exponentielle, on peut écrire que

$$P[T_2 > t \mid T_1 = \tau] = \begin{cases} e^{-\lambda_2 \tau} e^{-\lambda_3 (t-\tau)} & \text{si } 0 < \tau \le t \\ e^{-\lambda_2 t} & \text{si } \tau > t \end{cases}$$

Si la durée de vie du composant n° 2 est en fait indépendante de celle du composant n° 1, de sorte que $\lambda_3 = \lambda_2$, on obtient:

$$P[T_2 > t \mid T_1 = \tau] = \begin{cases} e^{-\lambda_2 \tau} e^{-\lambda_2 (t-\tau)} = e^{-\lambda_2 t} & \text{si } 0 < \tau \le t, \\ e^{-\lambda_2 t} & \text{si } \tau > t \end{cases}$$

C'est-à-dire que

$$P[T_2 > t \mid T_1 = \tau] = e^{-\lambda_2 t} \quad \text{pour } t \ge 0 \text{ et n'importe quel } \tau > 0$$

On a:

$$P[T_2 > t] = \int_0^\infty e^{-\lambda_2 t} f_{T_1}(\tau) d\tau = e^{-\lambda_2 t} \int_0^\infty f_{T_1}(\tau) d\tau = e^{-\lambda_2 t}$$

comme il se devait. ♢

Redondance passive

Supposons maintenant qu'un système comprend n sous-systèmes (numérotés de 1 à n) placés en parallèle, mais qu'un seul sous-système fonctionne à la fois. Au départ, seul le sous-système n° 1 est actif. Lorsqu'il tombe en panne, le sous-système n° 2 prend la relève, et ainsi de suite. Ce type de redondance est appelé *redondance passive* (ou *en attente*).

Remarques.
(i) Il est sous-entendu qu'il y a un dispositif qui envoie des signaux au système pour l'avertir d'activer le sous-système n° 2 lorsque le premier sous-système tombe en panne, et ainsi de suite. En pratique, ce dispositif lui-même peut tomber en panne. Cependant, nous supposons dans ce livre que le dispositif de signalisation demeure fiable à 100 % pendant un temps indéfini. Nous supposons aussi que les sous-systèmes placés en mode d'attente ne peuvent pas tomber en panne avant d'être activés, quoique nous puissions avoir deux distributions pour le temps écoulé jusqu'à une panne: une distribution pour un sous-système *dormant*, et une autre pour un sous-système *actif*.

(ii) Parce que les sous-systèmes fonctionnent l'un après l'autre, il est naturel de supposer (sauf indication contraire) que leurs durées de vie sont des variables aléatoires indépendantes.

La durée de vie totale T du système est évidemment donnée par

$$T = T_1 + T_2 + \cdots + T_n$$

de sorte que son temps moyen de fonctionnement jusqu'à une panne est

$$E[T] = \sum_{k=1}^{n} E[T_k]$$

De plus, puisque les sous-systèmes fonctionnent indépendamment les uns des autres, on a:

$$\mathrm{VAR}[T] \overset{\text{ind.}}{=} \sum_{k=1}^{n} \mathrm{VAR}[T_k]$$

En général, il n'est pas facile de trouver une expression explicite pour la fonction de fiabilité du système, parce que la fonction de densité de T est la convolution des fonctions de densité des variables aléatoires $T_1, \ldots, T_n$. Dans le cas particulier où $T_k \sim \mathrm{Exp}(\lambda)$, pour tout k, on sait (voir la sous-section 4.3.3) que T présente une distribution gamma de paramètres n et λ. De plus, en utilisant la formule (3.21), on peut écrire que

$$R(t) := P[T > t] = P[\mathrm{Poi}(\lambda t) \leq n-1] = \sum_{k=0}^{n-1} e^{-\lambda t} \frac{(\lambda t)^k}{k!} \quad \text{pour } t \geq 0$$

Exemple 8.2.4. Un système est constitué de deux composants identiques (et indépendants) placés en redondance passive. Si la durée de vie T_k de chaque composant présente une distribution uniforme sur l'intervalle $(0, 1)$, alors (voir l'exemple 4.3.4) on peut écrire que

$$F_T(t) = \begin{cases} \displaystyle\int_0^t s\, ds = \frac{t^2}{2} & \text{si } 0 < t < 1 \\ \displaystyle\frac{1}{2} + \int_1^t (2-s)\, ds = 2t - \frac{t^2}{2} - 1 & \text{si } 1 \leq t < 2 \end{cases}$$

Il s'ensuit que

$$R(t) = 1 - F_T(t) = \begin{cases} 1 - \dfrac{t^2}{2} & \text{si } 0 < t < 1 \\ 2 + \dfrac{t^2}{2} - 2t & \text{si } 1 \leq t < 2 \end{cases}$$

(et $R(t) = 0$ si $t \geq 2$). ♢

Exemple 8.2.5. Supposons que, dans l'exemple précédent, $T_1 \sim \mathrm{Exp}(\lambda_1)$ et $T_2 \sim \mathrm{Exp}(\lambda_2)$, où $\lambda_1 \neq \lambda_2$. Alors, en utilisant la formule (4.26), on a:

$$\begin{aligned} f_T(t) &= f_{T_1}(t_1) * f_{T_2}(t_2) = \int_{-\infty}^{\infty} f_{T_1}(u) f_{T_2}(t-u) du \\ &= \int_0^t \lambda_1 e^{-\lambda_1 u} \lambda_2 e^{-\lambda_2 (t-u)} du = e^{-\lambda_2 t} \int_0^t \lambda_1 \lambda_2 e^{-(\lambda_1 - \lambda_2) u} du \\ &\overset{\lambda_1 \neq \lambda_2}{=} \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2} \left(e^{-\lambda_2 t} - e^{-\lambda_1 t} \right) \quad \text{pour } t \geq 0 \end{aligned}$$

Il s'ensuit que

$$R(t) = \int_t^\infty f_T(s)ds = \frac{\lambda_1\lambda_2}{\lambda_1 - \lambda_2}\left(\frac{e^{-\lambda_2 t}}{\lambda_2} - \frac{e^{-\lambda_1 t}}{\lambda_1}\right)$$
$$= \frac{\lambda_1 e^{-\lambda_2 t} - \lambda_2 e^{-\lambda_1 t}}{\lambda_1 - \lambda_2} \quad \text{pour } t \geq 0$$

Notons que, en faisant appel à la règle de l'Hospital, on obtient:

$$\lim_{\lambda_2 \to \lambda_1} f_T(t) = \lambda_1^2 \lim_{\lambda_2 \to \lambda_1} \frac{e^{-\lambda_2 t}(-t) - 0}{0 - 1} = \lambda_1^2 t e^{-\lambda_1 t} \quad \text{pour } t \geq 0$$

C'est-à-dire que $T \sim \mathrm{G}(\alpha = 2, \lambda_1)$, comme il se devait. On a aussi:

$$\lim_{\lambda_2 \to \lambda_1} R(t) \overset{\text{L'Hos.}}{=} \lim_{\lambda_2 \to \lambda_1} \frac{\lambda_1 e^{-\lambda_2 t}(-t) - e^{-\lambda_1 t}}{0 - 1} = e^{-\lambda_1 t}(\lambda_1 t + 1) \quad \text{pour } t \geq 0$$

♢

8.2.3 Autres cas

Supposons qu'un système est formé de n sous-systèmes et qu'il doit y avoir au moins k sous-système(s) opérationnel(s) pour que celui-ci fonctionne, où $0 < k \leq n$. Ce type de système est appelé k *parmi* n. Notons qu'un système en série est le cas particulier où $k = n$, tandis qu'un système en parallèle (avec redondance active) correspond au cas où $k = 1$.

En général, on ne peut pas établir une formule simple pour la fonction de fiabilité $R(t)$ du système. Cependant, si tous les sous-systèmes sont indépendants et possèdent la même fonction de fiabilité $R_1(t)$, alors la fonction $R(t)$ est donnée par

$$R(t) = P[N \geq k], \quad \text{où } N \sim \mathrm{B}(n, p = R_1(t))$$

C'est-à-dire que

$$R(t) = \sum_{i=k}^{n} \binom{n}{i} [R_1(t)]^i [1 - R_1(t)]^{n-i} = 1 - \sum_{i=0}^{k-1} \binom{n}{i} [R_1(t)]^i [1 - R_1(t)]^{n-i}$$

Remarque. On suppose que les sous-systèmes fonctionnent indépendamment les uns des autres. En pratique, la durée de vie des composants qui fonctionnent dépend souvent du nombre total de composants actifs. Par exemple, supposons

qu'un avion possède quatre moteurs, mais qu'il peut voler et se poser avec seulement deux d'entre eux. Si deux moteurs tombent en panne pendant que l'avion est en vol, alors il y aura une plus grande charge sur les deux moteurs qui continuent à fonctionner, de sorte que leur durée de vie sera probablement plus courte.

Exemple 8.2.6. Considérons un système 2 parmi 3 pour lequel la durée de vie du sous-système i présente une distribution exponentielle de paramètre $\lambda_i = i$, pour $i = 1, 2, 3$. Pour obtenir la fonction de fiabilité du système, on utilise la formule suivante: si A_1, A_2 et A_3 sont des événements indépendants, alors

$$\begin{aligned} & P[(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3)] \\ &= P[A_1 \cap A_2] + P[A_1 \cap A_3] + P[A_2 \cap A_3] - 3P[A_1 \cap A_2 \cap A_3] \\ &\quad + P[A_1 \cap A_2 \cap A_3] \\ &\overset{\text{ind.}}{=} P[A_1]P[A_2] + P[A_1]P[A_3] + P[A_2]P[A_3] - 2P[A_1]P[A_2]P[A_3] \end{aligned}$$

Soit $A_i = \{T_i > t\}$, de sorte que

$$P[A_i] = P[T_i > t] = e^{-it} \quad \text{pour } i = 1, 2, 3$$

On peut alors écrire que

$$R(t) = e^{-3t} + e^{-4t} + e^{-5t} - 2e^{-6t} \quad \text{pour } t \geq 0$$

Remarque. On peut aussi écrire que

$$\begin{aligned} R(t) &= P[A_1 \cap A_2 \cap A_3'] + P[A_1 \cap A_2' \cap A_3] + P[A_1' \cap A_2 \cap A_3] \\ &\quad + P[A_1 \cap A_2 \cap A_3] \\ &\overset{\text{ind.}}{=} P[A_1]P[A_2](1 - P[A_3]) + P[A_1](1 - P[A_2])P[A_3] \\ &\quad + (1 - P[A_1])P[A_2]P[A_3] + P[A_1]P[A_2]P[A_3] \\ &= P[A_1]P[A_2] + P[A_1]P[A_3] + P[A_2]P[A_3] - 2P[A_1]P[A_2]P[A_3] \end{aligned}$$

comme ci-dessus. C'est-à-dire qu'on décompose l'événement $\{T > t\}$ en quatre cas incompatibles: *exactement* deux sous-systèmes fonctionnent à l'instant t, *ou* les trois sous-systèmes fonctionnent. ♢

Maintenant, le système présenté dans la figure 8.2 est appelé *système en pont*. Il fonctionne à l'instant t si et seulement si au moins un des événements suivants se produit:

- A_1: les composants n^{os} 1 et 4 sont actifs à l'instant t;
- A_2: les composants n^{os} 2 et 5 sont actifs à l'instant t;
- A_3: les composants n^{os} 1, 3 et 5 sont actifs à l'instant t;
- A_4: les composants n^{os} 2, 3 et 4 sont actifs à l'instant t.

Parce que les événements $A_1, \ldots, A_4$ ne sont ni indépendants ni incompatibles, on a besoin de la formule de la probabilité de l'union de quatre événements arbitraires:

$$\begin{aligned} P[A_1 \cup A_2 \cup A_3 \cup A_4] &= P[A_1] + P[A_2] + P[A_3] + P[A_4] - P[A_1 \cap A_2] \\ &\quad - P[A_1 \cap A_3] - P[A_1 \cap A_4] - P[A_2 \cap A_3] - P[A_2 \cap A_4] - P[A_3 \cap A_4] \\ &\quad + P[A_1 \cap A_2 \cap A_3] + P[A_1 \cap A_2 \cap A_4] + P[A_1 \cap A_3 \cap A_4] \\ &\quad + P[A_2 \cap A_3 \cap A_4] - P[A_1 \cap A_2 \cap A_3 \cap A_4] \end{aligned} \tag{8.1}$$

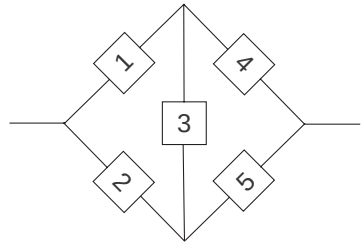

Fig. 8.2. Un système en pont

Dans le cas particulier où les cinq composants du système en pont fonctionnent indépendamment et possèdent tous la même fonction de fiabilité $R_1(t)$, on peut facilement calculer la fonction de fiabilité du système.

Finalement, comme nous l'avons fait au chapitre 2, on peut considérer des systèmes constitués de sous-systèmes placés en série et d'autres placés en parallèle.

Exemple 8.2.7. Un système est constitué de deux sous-systèmes placés en série. Le premier sous-système comprend deux composants connectés en parallèle, et le deuxième sous-système contient un seul composant. Supposons que les trois composants fonctionnent indépendamment. Soit $R_i(t)$ la fonction de fiabilité du

composant i, pour $i = 1, 2, 3$. Alors la fonction de fiabilité du système est donnée par

$$R(t) = \{1 - [1 - R_1(t)][1 - R_2(t)]\} R_3(t) = [R_1(t) + R_2(t) - R_1(t)R_2(t)] R_3(t)$$

C'est-à-dire qu'on utilise à la fois la formule des systèmes en série et celle des systèmes en parallèle. ♢

8.3 Liens et coupes

Lorsqu'un système est constitué d'un (relativement) grand nombre de composants, une première tâche en fiabilité consiste à trouver les divers ensembles de composants opérationnels qui vont permettre au système de fonctionner. Ces ensembles sont appelés *liens* (ou *chemins*). Inversement, on peut essayer de déterminer les ensembles de composants, appelés *coupes*, pour lesquels le système en entier cesse de fonctionner lorsque tous les composants qu'ils contiennent sont en panne.

Soit X_i une variable aléatoire de Bernoulli qui représente l'état du composant i à un instant fixé $t_0 \geq 0$. Plus précisément, $X_i = 1$ (respectivement $X_i = 0$) si le composant i est actif (respectivement en panne) à l'instant t_0, pour $i = 1, \ldots, n$. La variable aléatoire X_i est en fait la *variable indicatrice* de l'événement suivant: *le composant i est actif à l'instant t_0*. Pour calculer la fiabilité du système à l'instant t_0, on doit connaître la valeur de la probabilité $p_i := P[X_i]$, pour $i = 1, \ldots, n$. Cependant, pour déterminer les liens et les coupes d'un système, il suffit de considérer les valeurs particulières x_i prises par les variables aléatoires correspondantes. De plus, on suppose que le fait que le système fonctionne ou non à l'instant t_0 ne dépend que de l'état de ses composants à cet instant.

Définition 8.3.1. *La fonction*

$$H(x_1, \ldots, x_n) = \begin{cases} 1 \textit{ si le système fonctionne} \\ 0 \textit{ si le système est en panne} \end{cases}$$

est appelée **fonction de structure** *du système. Le vecteur* $\mathbf{x} := (x_1, \ldots, x_n)$ *est le* **vecteur d'état** *du système.*

Remarques.
(i) La fonction $H(\mathbf{X}) = H(X_1, \ldots, X_n)$ est elle-même une variable aléatoire de Bernoulli.

(ii) Soit $\mathbf{0} = (0, \ldots, 0)$ et $\mathbf{1} = (1, \ldots, 1)$. On suppose que $H(\mathbf{0}) = 0$ et $H(\mathbf{1}) = 1$. C'est-à-dire que si tous les composants du système sont en panne, alors celui-ci ne fonctionne plus et, inversement, si tous les composants sont actifs, alors le système fonctionne.

Notation. Considérons deux vecteurs en n dimensions: $\mathbf{x} = (x_1, \ldots, x_n)$ et $\mathbf{y} = (y_1, \ldots, y_n)$. On écrit:

$$\mathbf{x} \geq \mathbf{y} \quad \text{si } x_i \geq y_i \text{ pour } i = 1, \ldots, n$$

et

$$\mathbf{x} > \mathbf{y} \quad \text{si } x_i \geq y_i \text{ pour } i = 1, \ldots, n \text{ et } x_i > y_i \text{ pour au moins un } i$$

Définition 8.3.2. *Une fonction de structure $H(\mathbf{x})$ qui est telle que $H(\mathbf{0}) = 0$, $H(\mathbf{1}) = 1$ et*

$$H(\mathbf{x}) \geq H(\mathbf{y}) \quad \textit{si } \mathbf{x} \geq \mathbf{y} \tag{8.2}$$

est dite **monotone**.

Dans ce livre, nous supposons que les fonctions de structure $H(\mathbf{x})$ des systèmes considérés sont monotones.

Exemple 8.3.1. Dans le cas d'un système en série (constitué de n composants), on a:

$$H(\mathbf{x}) = 1 \quad \Longleftrightarrow \quad x_i = 1 \ \forall i \quad \left(\Longleftrightarrow \quad \sum_{i=1}^{n} x_i = n \right)$$

tandis que si les n composants sont placés en parallèle, alors

$$H(\mathbf{x}) = 1 \quad \Longleftrightarrow \quad \sum_{i=1}^{n} x_i \geq 1$$

En général, pour un système k parmi n, on peut écrire que

$$H(x_1, \ldots, x_n) = \begin{cases} 1 \text{ si } \sum_{i=1}^{n} x_i \geq k \\ 0 \text{ si } \sum_{i=1}^{n} x_i < k \end{cases}$$

Remarques.
(i) On peut aussi exprimer la fonction de structure $H(\mathbf{x})$ comme suit:

$$H(x_1, \ldots, x_n) = \begin{cases} \min\{x_1, \ldots, x_n\} \text{ pour un système en série} \\ \max\{x_1, \ldots, x_n\} \text{ pour un système en parallèle} \end{cases}$$

(ii) Dans cette section, lorsqu'on écrit que les composants sont placés en parallèle, on suppose qu'ils fonctionnent tous à partir de l'instant initial. C'est-à-dire qu'ils sont en redondance *active*. ♢

Pour calculer la valeur de la fonction de structure d'un système arbitraire, les formules suivantes sont utiles:

$$\min\{x_1, \ldots, x_n\} = \prod_{i=1}^{n} x_i$$

et

$$\max\{x_1, \ldots, x_n\} = 1 - \prod_{i=1}^{n}(1 - x_i)$$

Ces formules sont valables lorsque $x_i = 0$ ou 1, pour $i = 1, \ldots, n$.

Définition 8.3.3. *Un* **vecteur lien** *est n'importe quel vecteur* $\mathbf{x}$ *pour lequel* $H(\mathbf{x}) = 1$. *Si, en outre,* $H(\mathbf{y}) = 0$ *pour tous les vecteurs* $\mathbf{y}$ *tels que* $\mathbf{y} < \mathbf{x}$, *alors* $\mathbf{x}$ *est appelé* **vecteur lien minimal**. *De plus, à chaque vecteur lien minimal* $\mathbf{x} = (x_1, \ldots, x_n)$, *on associe un ensemble* $LM := \{k \in \{1, \ldots, n\} : x_k = 1\}$ *appelé* **(ensemble) lien minimal**.

Définition 8.3.4. *Si* $H(\mathbf{x}) = 0$, *on dit que le vecteur d'état* $\mathbf{x}$ *est un* **vecteur coupe**. *Si, en outre,* $H(\mathbf{y}) = 1$ *lorsque* $\mathbf{y} > \mathbf{x}$, *alors* $\mathbf{x}$ *est un* **vecteur coupe minimale**. *De plus, l'ensemble* $CM := \{k \in \{1, \ldots, n\} : x_k = 0\}$, *où* $\mathbf{x} = (x_1, \ldots, x_n)$ *est un vecteur coupe minimale, est appelé* **(ensemble) coupe minimale**.

Remarques.
(i) Dans certains livres, la définition d'un (ensemble) *lien* (respectivement *coupe*) correspond à celle que nous donnons de *lien minimal* (respectivement *coupe minimale*).

(ii) Un lien minimal est un groupe de composants tels que lorsqu'ils sont tous actifs, le système fonctionne, mais que si au moins un composant de ce groupe tombe en panne, alors le système aussi tombe en panne. Inversement, si tous les composants dans une coupe minimale sont en panne, le système est également en panne, mais si au moins un composant de la coupe minimale est remplacé par un composant opérationnel, alors le système fonctionne.

Exemple 8.3.2. Un système en série formé de n composants possède un seul lien minimal, soit l'ensemble $LM = \{1, 2, \ldots, n\}$ (parce que tous les composants

doivent fonctionner pour que le système fonctionne). Il possède n coupes minimales, lesquelles sont les ensembles contenant exactement un composant: $\{1\}$, $\ldots$, $\{n\}$. Notons que lorsqu'on écrit que $CM = \{1\}$, cela implique que les composants $2, \ldots, n$ sont opérationnels. De plus, le vecteur d'état $(0, 0, 1, 1, \ldots, 1)$ est un vecteur coupe, mais pas un vecteur coupe *minimale*, car si on remplace seulement le composant n° 1 par un composant opérationnel, le système demeure en panne.

Inversement, dans le cas d'un système en parallèle comprenant n composants, les liens minimaux sont $\{1\}$, $\ldots$, $\{n\}$, tandis qu'il n'y a qu'une seule coupe minimale: $\{1, 2, \ldots, n\}$. ♢

Exemple 8.3.3. On peut généraliser les résultats de l'exemple précédent comme suit: dans un système k parmi n, il y a $\binom{n}{k}$ liens minimaux. C'est-à-dire qu'on peut choisir n'importe quel ensemble de k composants parmi les n. Le nombre de coupes minimales est donné par $\binom{n}{n-k+1}$. En effet, s'il y a exactement $n - k + 1$ composants en panne, alors le système recommencera à fonctionner si l'un d'entre eux est remplacé par un composant opérationnel. ♢

Exemple 8.3.4. Le système en pont de la figure 8.2 possède quatre liens minimaux, comme mentionné indirectement ci-dessus: $\{1, 4\}$, $\{2, 5\}$, $\{1, 3, 5\}$ et $\{2, 3, 4\}$. Il possède également quatre coupes minimales: $\{1, 2\}$, $\{4, 5\}$, $\{1, 3, 5\}$ et $\{2, 3, 4\}$. Notons que $\{1, 3, 5\}$ et $\{2, 3, 4\}$ sont à la fois des liens minimaux et des coupes minimales. ♢

Supposons maintenant qu'un système quelconque possède r liens minimaux. Soit

$$\pi_j(x_1, \ldots, x_n) = \prod_{i \in LM_j} x_i \quad \text{pour } j = 1, \ldots, r$$

C'est-à-dire que $\pi_j(x_1, \ldots, x_n) = 1$ si tous les composants dans le lien minimal LM_j fonctionnent, et $\pi_j(x_1, \ldots, x_n) = 0$ autrement. Puisqu'un système fonctionne si et seulement si tous les composants dans au moins un de ses liens minimaux sont opérationnels, on peut représenter la fonction de structure du système en question comme suit:

$$H(\mathbf{x}) = 1 - \prod_{j=1}^{r} [1 - \pi_j(\mathbf{x})]$$

Cette formule implique qu'un système donné peut être considéré comme équivalent à celui obtenu en reliant ses liens minimaux en parallèle.

De la même manière, si un système quelconque possède s coupes minimales, on peut écrire que

$$H(x_1, \ldots, x_n) = 1 - \prod_{m=1}^{s} \gamma_m(x_1, \ldots, x_n)$$

où

$$\gamma_m(x_1, \ldots, x_n) := 1 - \prod_{i \in CM_m} (1 - x_i) \quad \text{pour } m = 1, \ldots, s$$

On a que $\gamma_m(x_1, \ldots, x_n)$ est égal à 0 si tous les composants de la coupe minimale CM_m sont en panne, et à 1 autrement. Cette fois-ci, on peut affirmer qu'un système donné et celui formé de ses coupes minimales reliées en série sont équivalents.

Exemple 8.3.5. À partir de l'exemple précédent, on déduit que le système en pont de la figure 8.2 est équivalent au système présenté dans la figure 8.3 ou à celui de la figure 8.4. ♢

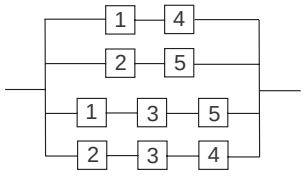

Fig. 8.3. Un système en pont représenté comme un système en parallèle formé de ses liens minimaux

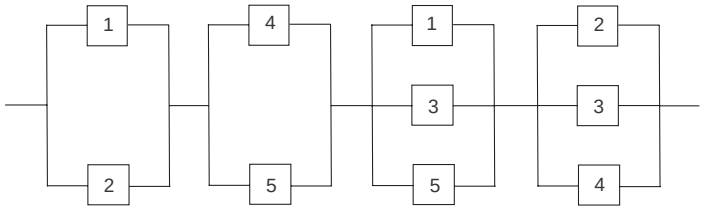

Fig. 8.4. Un système en pont représenté comme un système en série formé de ses coupes minimales

Parce que $H(\mathbf{X})$, où $\mathbf{X} := (X_1, \ldots, X_n)$, est une variable aléatoire de Bernoulli, la fiabilité du système, à l'instant fixé t_0, est donnée par

$$R(t_0) = P[H(\mathbf{X}) = 1] = E[H(\mathbf{X})]$$

Si l'on pose:

$$p_i = P[X_i = 1] \quad (\text{à l'instant } t_0) \quad \text{pour } i = 1, \ldots, n$$

et si l'on suppose que les composants fonctionnent indépendamment les uns des autres, alors on peut écrire que

$$R(t_0) = \begin{cases} \prod_{i=1}^{n} p_i & \text{pour un système en série} \\ 1 - \prod_{i=1}^{n} (1 - p_i) & \text{pour un système en parallèle} \end{cases}$$

De plus, si $p_i = p$ pour tout i, alors

$$R(t_0) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

dans le cas d'un système k parmi n. Ces formules sont simplement des cas particuliers des formules correspondantes de la section 8.2.

Finalement, lorsqu'un système comprend plusieurs composants, on peut au moins essayer d'obtenir des bornes pour sa fiabilité $R(t_0)$ à l'instant t_0. On peut montrer que

$$\prod_{m=1}^{s} P[\gamma_m(\mathbf{X}) = 1] \leq R(t_0) \leq 1 - \prod_{j=1}^{r} \{1 - P[\pi_j(\mathbf{X}) = 1]\}$$

où

$$P[\gamma_m(\mathbf{X}) = 1] = 1 - \prod_{i \in CM_m} (1 - p_i)$$

et

$$P[\pi_j(\mathbf{X}) = 1] = \prod_{i \in LM_j} p_i$$

Exemple 8.3.6. Supposons que $p_i \equiv 0{,}9$ pour le système en pont de la figure 8.2. En utilisant l'équation (8.1), on trouve (en supposant que les composants sont indépendants) que

$$R(t_0) = 2(0{,}9)^2 + 2(0{,}9)^3 - 5(0{,}9)^4 + 2(0{,}9)^5 = 0{,}97848$$

En effet, on a que $P[A_i] = (0{,}9)^2$, pour $i = 1, 2$, et $P[A_i] = (0{,}9)^3$, pour $i = 3, 4$. De plus, la probabilité de n'importe quelle intersection dans l'équation (8.1) est égale à $(0{,}9)^k$, où k est le nombre de composants distincts impliqués dans l'intersection en question. Par exemple, $A_1 \cap A_2$ se produit si et seulement si les composants 1, 2, 4 et 5 sont opérationnels, de sorte que $P[A_1 \cap A_2] = (0{,}9)^4$.

Maintenant, on déduit de l'exemple 8.3.4 que la borne inférieure pour la probabilité $R(t_0)$ est donnée par

$$\left[1 - (0{,}1)^2\right]^2 \left[1 - (0{,}1)^3\right]^2 \simeq 0{,}97814$$

et que la borne supérieure est

$$1 - \left\{\left[1 - (0{,}9)^2\right]^2 \left[1 - (0{,}9)^3\right]^2\right\} \simeq 0{,}99735$$

Notons que dans cet exemple la borne inférieure, en particulier, est très précise. ♢

8.4 Exercices du chapitre 8

Exercices résolus

Question n^o 1

On suppose que la durée de vie X d'un certain système peut être exprimée comme suit: $X = Y^2$, où Y est une variable aléatoire distribuée uniformément sur l'intervalle $(0, 1)$. Trouver la fonction de fiabilité du système.

Solution. On a:

$$R(x) := P[X > x] = P[Y^2 > x] = \begin{cases} P[Y > \sqrt{x}] = 1 - \sqrt{x} & \text{si } 0 \leq x < 1 \\ 0 & \text{si } x \geq 1 \end{cases}$$

Question n^o 2

Le temps (en années) écoulé jusqu'à une panne pour un dispositif donné est une variable aléatoire X qui présente une distribution exponentielle de paramètre $\lambda = 1/2$. Quand le dispositif tombe en panne, il est réparé. Le temps de réparation Y (en jours) est une variable aléatoire telle que

$$P[Y > y] = \begin{cases} e^{-y} & \text{si } X < 2 \text{ et } y \geq 0 \\ e^{-y/2} & \text{si } X \geq 2 \text{ et } y \geq 0 \end{cases}$$

Calculer le temps moyen entre deux pannes (soit la quantité $MTBF$) pour ce dispositif.

Solution. On peut écrire que $Y \mid \{X < 2\} \sim \text{Exp}(1)$ et $Y \mid \{X \geq 2\} \sim \text{Exp}(1/2)$. Il s'ensuit que

$$\begin{aligned} E[Y] &= E[Y \mid X < 2]P[X < 2] + E[Y \mid X \geq 2]P[X \geq 2] \\ &= 1 \times \left[1 - e^{-(1/2)2}\right] + 2 \times e^{-(1/2)2} = 1 + e^{-1} \end{aligned}$$

Puisque $E[X] = 2$ ans, on a que $MTBF = 731 + e^{-1} \simeq 731{,}37$ jours.

Question n$^{\text{o}}$ 3

Soit T une variable aléatoire continue qui possède la fonction de densité de probabilité suivante:

$$f_T(t) = \frac{1}{\lambda} \exp\left\{-\frac{e^t - 1}{\lambda} + t\right\} \quad \text{si } t \geq 0$$

(et $f_T(t) = 0$ si $t < 0$), où λ est un paramètre positif. On dit que T présente une *distribution des valeurs extrêmes* (particulière). Calculer le taux de panne pour un dispositif dont la durée de vie est distribuée comme T.

Solution. D'abord, on calcule

$$\begin{aligned} F_T(t) &= \int_0^t f_T(s)ds = \int_0^t \frac{1}{\lambda} \exp\left\{-\frac{e^s - 1}{\lambda} + s\right\} ds \\ &= \int_0^t \frac{1}{\lambda} e^s \exp\left\{-\frac{e^s - 1}{\lambda}\right\} ds = -\exp\left\{-\frac{e^s - 1}{\lambda}\right\}\Bigg|_0^t \\ &= 1 - \exp\left\{-\frac{e^t - 1}{\lambda}\right\} \quad \text{pour } t \geq 0 \end{aligned}$$

Il s'ensuit que

$$r(t) = \frac{f_T(t)}{1 - F_T(t)} = \frac{e^t}{\lambda} \quad \text{pour } t \geq 0$$

Question n$^{\text{o}}$ 4

On suppose que la durée de vie (en cycles) d'un système est une variable aléatoire de Poisson de paramètre $\lambda > 0$. Le taux de panne $r(k)$ est-il croissant ou décroissant en $k = 0$?

Solution. On calcule

$$r(0) = \frac{e^{-\lambda}}{\sum_{j=0}^{\infty} e^{-\lambda}\lambda^j/j!} = \frac{e^{-\lambda}}{1} = e^{-\lambda}$$

et

$$r(1) = \frac{e^{-\lambda}\lambda}{\sum_{j=1}^{\infty} e^{-\lambda}\lambda^j/j!} = \frac{e^{-\lambda}\lambda}{e^{-\lambda}(e^{\lambda}-1)} = \frac{\lambda}{e^{\lambda}-1}$$

On a:

$$r(0) < r(1) \quad \Longleftrightarrow \quad e^{-\lambda} < \frac{\lambda}{e^{\lambda}-1} \quad \Longleftrightarrow \quad 1 - e^{-\lambda} < \lambda$$

Soit

$$g(\lambda) = \lambda + e^{-\lambda} - 1$$

Puisque $g(0) = 0$ et

$$g'(\lambda) = 1 - e^{-\lambda} > 0 \quad \text{pour tout } \lambda > 0$$

on peut affirmer que $g(\lambda) > 0$, pour $\lambda > 0$. De là, on conclut que $r(0) < r(1)$, de sorte que le taux de panne est croissant en $k = 0$.

Remarque. En fait, on peut montrer que la fonction $r(k)$ est croissante en n'importe quel $k \in \{0, 1, \ldots\}$.

Question n$^{\text{o}}$ 5

Si un système a une durée de vie X qui est distribuée uniformément sur l'intervalle $[0, 1]$, quel est le taux moyen de panne (AFR) dans l'intervalle $[0, 1/2]$?

Solution. On a:

$$R(x) := P[X > x] = \int_x^1 1\,dt = 1 - x \quad \text{si } 0 \le x \le 1$$

Par conséquent,

$$AFR(0, 1/2) = \frac{\ln[R(0)] - \ln[R(1/2)]}{1 - (1/2)} = 2[\ln 1 - \ln(1/2)] = 2\ln 2$$

Question n$^{\text{o}}$ 6

Un système comprend deux composants (indépendants) placés en série. La durée de vie X_k du composant k présente une distribution exponentielle de paramètre λ_k, pour $k = 1, 2$. Utiliser la formule (voir l'équation (4.15))

$$P[X_2 < X_1] = \int_{-\infty}^{\infty} P[X_2 < x_1 \mid X_1 = x_1] f_{X_1}(x_1) dx_1$$

où X_1 et X_2 sont des variables aléatoires continues quelconques, pour calculer la probabilité que la première panne du système soit causée par une panne du composant n$^{\text{o}}$ 2.

Solution. Les composants sont placés en série; par conséquent, on cherche effectivement la probabilité $P[X_2 < X_1]$. En utilisant la formule ci-dessus, on peut écrire que

$$P[X_2 < X_1] = \int_{0}^{\infty} P[X_2 < x_1 \mid X_1 = x_1] \lambda_1 e^{-\lambda_1 x_1} dx_1$$

Par indépendance, on obtient que

$$\begin{aligned} P[X_2 < X_1] &= \int_0^{\infty} [1 - e^{-\lambda_2 x_1}] \lambda_1 e^{-\lambda_1 x_1} dx_1 = 1 - \lambda_1 \int_0^{\infty} e^{-(\lambda_1+\lambda_2)x_1} dx_1 \\ &= 1 - \lambda_1 \left(\frac{1}{\lambda_1 + \lambda_2} \right) = \frac{\lambda_2}{\lambda_1 + \lambda_2} \end{aligned}$$

Remarque. Notons que $P[X_2 < X_1] = 1/2$ si $\lambda_1 = \lambda_2$, ce qui est logique, par symétrie (car $P[X_2 = X_1] = 0$, par continuité).

Question n$^{\text{o}}$ 7

On considère un système constitué de deux composants placés en parallèle et fonctionnant indépendamment l'un de l'autre. Soit T la durée de vie totale du système, et soit T_k la durée de vie du composant k, pour $k = 1, 2$. Supposons que $T_k \sim \text{Exp}(\lambda_k)$. Quelle est la probabilité que les deux composants soient encore actifs à l'instant $t_0 > 0$, étant donné que le système fonctionne à cet instant?

Solution. On a que $P[T > t_0] = P[\{T_1 > t_0\} \cup \{T_2 > t_0\}]$. Soit $A_k = \{T_k > t_0\}$, pour $k = 1, 2$. On cherche

$$p := P[A_1 \cap A_2 \mid A_1 \cup A_2] = \frac{P[A_1 \cap A_2]}{P[A_1 \cup A_2]}$$

car $\{A_1 \cap A_2\} \subset \{A_1 \cup A_2\}$.

Ensuite, on a:

$$P[A_1 \cap A_2] \overset{\text{ind.}}{=} P[A_1] P[A_2] = e^{-(\lambda_1+\lambda_2)t_0}$$

et

$$P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2] = e^{-\lambda_1 t_0} + e^{-\lambda_2 t_0} - e^{-(\lambda_1+\lambda_2)t_0}$$

de sorte que

$$p = \frac{e^{-(\lambda_1+\lambda_2)t_0}}{e^{-\lambda_1 t_0} + e^{-\lambda_2 t_0} - e^{-(\lambda_1+\lambda_2)t_0}}$$

Question n^o 8

On dispose de deux composants identiques de marque A et de deux composants identiques de marque B. Pour construire un certain dispositif, on doit relier un composant de marque A et un composant de marque B en série. On suppose que la fiabilité de chaque composant est égale à 0,9 (à l'instant initial) et qu'ils fonctionnent tous indépendamment les uns des autres. Est-il préférable de construire deux dispositifs distincts et d'espérer qu'au moins un des deux fonctionnera, ou bien de construire un dispositif constitué de deux sous-systèmes connectés en série, le premier (respectivement deuxième) sous-système comportant les deux composants de marque A (respectivement B) placés en parallèle?

Solution. Soit F: le dispositif fonctionne à l'instant initial. Dans le cas d'un dispositif constitué d'un composant de marque A et d'un composant de marque B placés en série, on a que $P[F] = (0{,}9)^2 = 0{,}81$. Par conséquent, la probabilité qu'au moins un des deux dispositifs fonctionne est donnée par $1 - (1 - 0{,}81)^2 = 0{,}9639$.

Si l'on construit un seul dispositif comme décrit ci-dessus, on a:

$$P[F] = (1 - (0{,}1)^2)(1 - (0{,}1)^2) = 0{,}9801$$

Donc, il est préférable de doubler les composants plutôt que de doubler les dispositifs.

Remarque. La conclusion de cet exercice peut être généralisée comme suit: il est toujours préférable de doubler les composants dans un système en série plutôt que de construire deux systèmes distincts.

Question n^o 9

Calculer la fonction de structure du système représenté dans la figure 2.14, page 61, en fonction des variables indicatrices x_k, pour $k = 1, \ldots, 4$.

Solution. On a:

$$\begin{aligned} H(x_1, x_2, x_3, x_4) &= \max\{x_1x_2, x_3\}x_4 = [1 - (1 - x_1x_2)(1 - x_3)]x_4 \\ &= (x_1x_2 + x_3 - x_1x_2x_3)x_4 \end{aligned}$$

Question n$^{\text{o}}$ 10

Trouver les liens minimaux du système représenté dans la figure 2.14, page 61, et exprimer la fonction de structure $H(x_1, x_2, x_3, x_4)$ à l'aide des fonctions $\pi_j(x_1, x_2, x_3, x_4)$.

Solution. Les liens minimaux du système sont les suivants: $LM_1 = \{1, 2, 4\}$ et $LM_2 = \{3, 4\}$. De là,

$$\pi_1(x_1, x_2, x_3, x_4) = x_1x_2x_4 \quad \text{et} \quad \pi_2(x_1, x_2, x_3, x_4) = x_3x_4$$

de sorte que

$$H(x_1, x_2, x_3, x_4) = 1 - (1 - x_1x_2x_4)(1 - x_3x_4)$$

Remarque. Puisque $x_k = 0$ ou 1 $\forall k$, on peut écrire que $x_k^2 = x_k$. Il s'ensuit que

$$\begin{aligned} H(x_1, \ldots, x_4) &= x_1x_2x_4 + x_3x_4 - x_1x_2x_3x_4^2 = x_1x_2x_4 + x_3x_4 - x_1x_2x_3x_4 \\ &= (x_1x_2 + x_3 - x_1x_2x_3)x_4 \end{aligned}$$

ce qui est en accord avec le résultat de l'exercice précédent.

Exercices

Question n$^{\text{o}}$ 1

On veut construire un système constitué de deux composants placés en parallèle, suivis d'un composant placé en série. On suppose qu'on a trois composants et que tout arrangement des composants à l'intérieur du système est permis. Si la fiabilité du composant n$^{\text{o}}$ k, à un instant fixé $t_0 > 0$, est égale à p_k, pour $k = 1, 2, 3$, et si $0 < p_1 < p_2 < p_3 < 1$, quel arrangement des trois composants donne la plus grande probabilité que le système fonctionne à l'instant t_0?

Question n$^{\text{o}}$ 2

La durée de vie X d'une certaine machine possède la fonction de densité de probabilité suivante:

$$f_X(x) = \begin{cases} 1/x \text{ si } 1 \leq x \leq e \\ 0 \quad \text{ailleurs} \end{cases}$$

Calculer le taux de panne $r(x)$, pour $1 \leq x \leq e$. La distribution de X est-elle IFR ou DFR? Justifier la réponse.

Question n° 3

On suppose que le taux de panne $r(t)$ d'un système donné est $r(t) = 1$ $\forall t \geq 0$. Trouver la probabilité que le système tombe en panne dans l'intervalle $[2, 3]$, étant donné qu'il fonctionne encore à l'instant $t = 1$.

Question n° 4

La durée de vie T d'un appareil présente une distribution exponentielle de paramètre Λ, où Λ est une variable aléatoire distribuée uniformément sur l'intervalle $[1, 3]$. Calculer la fonction de fiabilité de cet appareil.

Question n° 5

Soit

$$f_T(t) = \begin{cases} \frac{1}{2}t^2 e^{-t} \text{ si } t > 0 \\ 0 \quad \text{ailleurs} \end{cases}$$

où T symbolise la durée de vie d'un certain système. Calculer le taux de panne $FR(0, 1)$ de ce système dans l'intervalle $(0, 1]$.

Question n° 6

On suppose que la durée de vie T d'un système présente une distribution G$(4, 1)$; c'est-à-dire que

$$f_T(t) = \frac{1}{6}t^3 e^{-t} \quad \text{pour } t \geq 0$$

Calculer la fonction de fiabilité du système en question à l'instant $t = 4$, étant donné qu'il fonctionne encore à l'instant $t = 2$.

Question n° 7

Calculer le taux de panne à l'instant $t = 1$ d'un système dont la durée de vie T est distribuée comme $|Z|$, où $Z \sim \text{N}(0, 1)$.

Question n° 8

Soit

$$p_X(k) = \frac{1}{N} \quad \text{pour } k = 1, 2, \ldots, N$$

la fonction de probabilité de la durée de vie X (en cycles) d'un système particulier. Calculer $r_X(k)$, pour $k = 1, \ldots, N$. Est-ce que X possède une distribution IFR ou une distribution DFR? Justifier la réponse.

Question n° 9

On considère le taux de panne d'un certain système dans l'intervalle $(n, n+1]$, pour $n = 0, 1, \ldots, 9$. Calculer le taux moyen de panne de ce système dans $(n, n+1]$, pour $n = 0, 1, \ldots, 9$, si $T \sim \mathrm{U}[0, 10]$.

Question n° 10

On suppose que T présente une distribution de Weibull de paramètres $\lambda > 0$ et $\beta > 0$. Quel est le taux moyen de panne dans l'intervalle $[0, \tau]$, si τ est une variable aléatoire distribuée comme la racine carrée d'une distribution $\mathrm{U}(0, 1)$ et (a) $\beta = 2$? (b) $\beta = 3$?

Question n° 11

Trois composants indépendants sont connectés en parallèle. On suppose que la durée de vie du composant n° k présente une distribution exponentielle de paramètre λ_k, pour $k = 1, 2, 3$. Quelle est la probabilité que le composant n° 3 ne soit pas le premier à tomber en panne?

Question n° 12

Un système comporte deux composants qui fonctionnent indépendamment l'un de l'autre, à partir de l'instant initial pour chacun d'eux. On suppose que la durée de vie X_k (en cycles) du composant n° k présente une distribution géométrique de paramètre $p_k = 1/2$, pour $k = 1, 2$. Trouver la probabilité que les deux composants tombent en panne pendant le même cycle.

Question n° 13

Trois composants indépendants sont placés en série. On suppose que la durée de vie T_k du k^{e} composant présente une distribution uniforme sur l'intervalle $(0, k+1)$, pour $k = 1, 2, 3$. Quelle est la valeur de la fonction de fiabilité du système (constitué de ces trois composants) à l'instant $t = 1$, étant donné qu'au moins un des trois composants n'est pas en panne à cet instant?

Question n° 14

On considère le système représenté dans la figure 2.13, page 60. On suppose que les composants A ont une durée de vie qui présente une distribution exponentielle de paramètre λ_A, et que la durée de vie du composant B (respectivement C) présente une distribution exponentielle de paramètre λ_B (respectivement λ_C). Calculer la fonction de fiabilité du système, en supposant que les composants fonctionnent indépendamment les uns des autres.

Question n° 15

Un système est constitué de deux composants qui fonctionnent indépendamment l'un de l'autre et qui sont placés en parallèle. On suppose que la durée de

vie T_k du k^{e} composant présente une distribution exponentielle de paramètre k, pour $k = 1, 2$. Quelle est la probabilité que le système fonctionne encore à l'instant $t = 2$, étant donné qu'exactement un de ses composants est en panne à l'instant $t = 1$?

Question n$^{\text{o}}$ 16

On dispose de quatre composants indépendants ayant une durée de vie qui présente une distribution exponentielle. L'espérance mathématique de la durée de vie du k^{e} composant est égale à $1/k$, pour $k = 1, 2, 3, 4$. Les composants sont utilisés pour construire un système constitué de deux sous-systèmes placés en série. Chaque sous-système comporte deux composants placés en parallèle. Quelle est l'espérance mathématique de la durée de vie du système, si le premier sous-système contient le premier et le deuxième composant (voir la figure 8.5)?

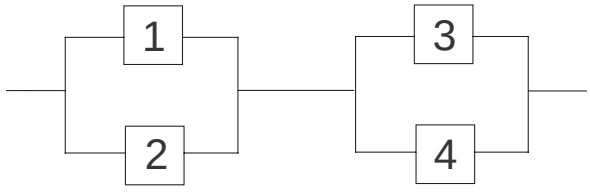

Fig. 8.5. Figure pour l'exercice n$^{\text{o}}$ 16

Question n$^{\text{o}}$ 17

On numérote les composants du système de la figure 2.13, page 60, comme suit: le composant $C = 1$, les composants $A = 2$, 3 et 4, et le composant $B = 5$. Trouver les liens minimaux et les coupes minimales de ce système.

Question n$^{\text{o}}$ 18

On suppose que la fiabilité des composants (indépendants) de la question précédente est la même que dans l'exercice non résolu n$^{\text{o}}$ 8 du chapitre 2, page 60 (à un instant fixé $t_0 > 0$). Utiliser les liens minimaux et les coupes minimales trouvés dans la question précédente pour calculer les bornes inférieure et supérieure pour la fiabilité du système. Comparer ces bornes à la réponse exacte.

Question n^{o} 19

Un certain système constitué de quatre composants indépendants fonctionne si et seulement si le composant n^{o} 1 est opérationnel et au moins deux des trois autres composants fonctionnent. Quels sont les liens minimaux et les coupes minimales de ce système?

Question n^{o} 20

Dans la question précédente,

(a) quelle est la probabilité que le système fonctionne à l'instant t_0, si la fiabilité de chaque composant est égale à p à cet instant?

(b) quelle est la fonction de fiabilité du système à l'instant t, si la durée de vie de chaque composant est une variable aléatoire Exp(θ)?

Questions à choix multiple

Question n^{o} 1

On suppose que $X \sim \text{Exp}(1)$ et $Y \sim \text{U}(0,1)$ sont des variables aléatoires indépendantes. On définit $Z = \min\{X, Y\}$. Trouver le taux de panne $r_Z(t)$ de Z pour $0 < t < 1$.

(a) 1 (b) 2 (c) $\dfrac{1}{1-t}$ (d) $\dfrac{2-t}{1-t}$ (e) $\dfrac{2}{1-t}$

Question n^{o} 2

La durée de vie T d'un dispositif présente une distribution lognormale de paramètres $\mu = 10$ et $\sigma^2 = 4$. C'est-à-dire que $\ln T \sim \text{N}(10, 4)$. Trouver la fiabilité du dispositif à l'instant $t = 200$.

(a) 0,01 (b) 0,05 (c) 0,5 (d) 0,95 (e) 0,99

Question n^{o} 3

Pour quelles valeurs du paramètre α la distribution bêta de paramètres α (> 0) et $\beta = 1$ est-elle une distribution IFR partout dans l'intervalle $(0, 1)$?

(a) n'importe quel $\alpha > 0$ (b) aucun α (c) $\alpha \geq 1$ (d) $\alpha < 1$ (e) $\alpha \geq 2$

Question n^{o} 4

On suppose que la durée de vie T d'un système est distribuée uniformément sur l'intervalle $(0, B)$, où B est une variable aléatoire présentant une distribution exponentielle de paramètre $\lambda > 0$. Trouver la fonction de fiabilité du système.

(a) $\dfrac{1}{\lambda} - t$ (b) $\lambda - t$ (c) $1 - \dfrac{t}{\lambda}$ (d) $1 - \lambda t$ (e) $\lambda(1-t)$

Question n° 5

Calculer le taux moyen de panne dans l'intervalle $[0, 2]$ d'un système dont la durée de vie T possède la fonction de densité de probabilité suivante:

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

où λ est un paramètre positif.

(a) $\dfrac{\lambda - 1}{2}$ (b) $\dfrac{\lambda}{2}$ (c) $\lambda - 1$ (d) λ (e) 2λ

Question n° 6

Deux composants indépendants sont connectés en série. La durée de vie T_k du composant n° k présente une distribution exponentielle de paramètre λ_k, pour $k = 1, 2$. Quand un composant tombe en panne, il est réparé. On suppose que la durée de vie T_k^* d'un composant réparé présente aussi une distribution exponentielle, mais est deux fois plus courte, en moyenne, que celle d'un composant neuf, de sorte que $T_k^* \sim \text{Exp}(2\lambda_k)$, pour $k = 1, 2$. Trouver la probabilité que le composant n° 1 cause les deux premières pannes du système en série.

(a) $\dfrac{\lambda_1^2}{(\lambda_1 + \lambda_2)^2}$ (b) $\dfrac{\lambda_1^2}{(\lambda_1 + 2\lambda_2)^2}$ (c) $\dfrac{\lambda_1^2}{(\lambda_1 + 2\lambda_2)(\lambda_1 + \lambda_2)}$

(d) $\dfrac{\lambda_1^2}{(2\lambda_1 + \lambda_2)(\lambda_1 + \lambda_2)}$ (e) $\dfrac{\lambda_1^2}{(2\lambda_1 + \lambda_2)^2}$

Question n° 7

On considère un système constitué de deux composants indépendants placés en redondance passive. La durée de vie T_1 du composant n° 1 présente une distribution uniforme sur l'intervalle $(0, 2)$, tandis que la durée de vie T_2 du composant n° 2 présente une distribution exponentielle de paramètre $\lambda = 2$. De plus, on suppose que le composant n° 2 prend la relève du premier dès que celui-ci tombe en panne si $T_1 < 1$, ou à l'instant $t = 1$ si le composant n° 1 est encore opérationnel à cet instant. Quelle est la durée de vie moyenne du système?

(a) 9/4 (b) 5/2 (c) 11/4 (d) 3 (e) 13/4

Question n° 8

Un certain système 15 parmi 20 est tel que tous ses composants (indépendants) ont une probabilité de 3/4 d'être opérationnels à l'instant $t_0 > 0$.

(i) Calculer la probabilité p que le système fonctionne à $t = t_0$.

(ii) Utiliser une approximation de Poisson pour calculer la probabilité p en (i).

(a) (i) 0,4148; (ii) 0,4405 (b) (i) 0,6172; (ii) 0,5343 (c) (i) 0,6172; (ii) 0,6160
(d) (i) 0,7858; (ii) 0,6380 (e) (i) 0,7858; (ii) 0,7622

Question n^o 9

On considère le système de la figure 8.6. Combien (i) de liens minimaux et (ii) de coupes minimales possède-t-il?

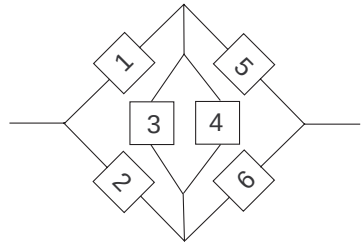

Fig. 8.6. Figure pour la question à choix multiple n^o 9

(a) (i) 4; (ii) 4 (b) (i) 5; (ii) 4 (c) (i) 5; (ii) 5 (d) (i) 6; (ii) 4
(e) (i) 6; (ii) 5

Question n^o 10

On suppose que, dans la question précédente, les composants fonctionnent indépendamment les uns des autres et ont tous une probabilité de 3/4 d'être opérationnels à l'instant $t_0 > 0$. Donner une borne inférieure pour la fiabilité du système à l'instant t_0.

(a) 0,1279 (b) 0,3164 (c) 0,6836 (d) 0,8721 (e) 0,8789

9

Files d'attente

Une application importante des probabilités est le domaine connu sous le nom de *théorie des files d'attente*. Ce domaine étudie le comportement de files d'attente ou *queues*. Les ingénieurs en télécommunication et les informaticiens sont particulièrement intéressés à la théorie des files d'attente pour résoudre des problèmes d'allocation et d'utilisation efficace des ressources dans des réseaux sans fil et des réseaux d'ordinateurs, par exemple. En général, les modèles considérés dans ce chapitre sont tels que les *arrivées* dans le *système de file d'attente* et les *départs* de ce système constituent deux processus de Poisson, que nous avons définis au chapitre 3 (page 86). Les processus de Poisson sont en fait des *chaînes de Markov à temps continu* particulières, lesquelles sont le sujet de la première section de ce chapitre. Le cas où il y a un seul *serveur* dans le système de file d'attente et celui où le nombre de serveurs est supérieur ou égal à deux sont étudiés séparément.

9.1 Chaînes de Markov à temps continu

Définition 9.1.1. *Un* **processus stochastique** *(ou* **processus aléatoire***) est un ensemble* $\{X(t), t \in T\}$ *de variables aléatoires* $X(t)$*, où* T *est un sous-ensemble de* $\mathbb{R}$*.*

Remarques.
(i) La variable *déterministe* t est souvent interprétée comme étant le *temps* dans les problèmes considérés. Dans ce chapitre, nous nous intéressons aux processus stochastiques *à temps continu*, de sorte que l'ensemble T est généralement l'intervalle $[0, \infty)$.

(ii) L'ensemble de toutes les valeurs possibles des variables aléatoires $X(t)$ est appelé *espace des états* du processus stochastique.

Une classe très importante de processus stochastiques pour les applications est celle des processus connus sous le nom de *processus de Markov*.

Définition 9.1.2. *Si on peut écrire que*

$$P[X(t_n) \leq x_n \mid X(t), \forall t \leq t_{n-1}] = P[X(t_n) \leq x_n \mid X(t_{n-1})] \tag{9.1}$$

où $t_{n-1} < t_n$, *on dit que le processus stochastique* $\{X(t), t \in T\}$ *est un* **processus de Markov** *(ou* **processus markovien***).*

Remarque. L'équation précédente, appelée *propriété de Markov*, signifie que le *futur* du processus ne dépend que de son *état présent*. C'est-à-dire que, en supposant que l'ensemble T est l'intervalle $[0, \infty)$, on n'a pas besoin de connaître l'*histoire* du processus dans l'intervalle $[0, t_{n-1})$ pour calculer la distribution de la variable aléatoire $X(t_n)$, où $t_n > t_{n-1}$, *si* la valeur de $X(t_{n-1})$ est connue.

Définition 9.1.3. *Si on suppose que l'ensemble des valeurs que les diverses variables aléatoires* $X(t)$ *peuvent prendre est au plus infini dénombrable, de sorte que* $X(t)$ *est une variable aléatoire discrète pour n'importe quelle valeur fixée de la variable* t, *alors on dit que* $\{X(t), t \in T\}$ *est un processus stochastique* **à état discret**.

Maintenant, soit τ_i le temps que le processus markovien à temps continu et à état discret $\{X(t), t \geq 0\}$ passe dans un état donné i avant de faire une transition vers n'importe quel autre état. On déduit de la propriété de Markov que

$$P[\tau_i > s + t \mid \tau_i > s] = P[\tau_i > t] \quad \forall s, t \geq 0 \tag{9.2}$$

(autrement le *futur* dépendrait du *passé*). Cette équation implique que la variable aléatoire continue τ_i présente une distribution exponentielle. En effet, seule la distribution exponentielle possède la *propriété de non-vieillissement* (voir la page 94) en temps continu.

Remarques.

(i) On symbolise le paramètre de la variable aléatoire τ_i par ν_i, pour tout i. Dans le cas général, ν_i dépend de l'état correspondant i. Cependant, dans le cas d'un processus de Poisson de taux λ, on a que $\nu_i = \lambda$ pour tout i.

(ii) On déduit aussi de la propriété de Markov que l'état qui sera visité lorsque le processus quittera l'état courant i doit être *indépendant* du temps total τ_i que le processus aura passé dans i avant de faire une transition.

Définition 9.1.4. *Le processus stochastique à temps continu et à état discret* $\{X(t), t \geq 0\}$ *est appelé* **chaîne de Markov à temps continu** *si*

$$P[X(t) = j \mid X(s) = i, X(r) = x_r, 0 \leq r < s] = P[X(t) = j \mid X(s) = i]$$

pour tout $t \geq s$ *et pour tous les états* i, j, x_r.

Remarques.
(i) On suppose que les chaînes de Markov considérées possèdent des probabilités de transition *homogènes par rapport au temps.* Cela signifie que si $t \geq s \geq 0$ et $\tau \geq 0$, on peut écrire que

$$P[X(t) = j \mid X(s) = i] := p_{i,j}(t - s) \iff P[X(\tau + t) = j \mid X(\tau) = i] = p_{i,j}(t)$$

C'est-à-dire que la probabilité que le processus passe de l'état i à l'état j dans un intervalle de temps donné ne dépend que de la longueur de cet intervalle de temps. Cette hypothèse est faite dans la plupart des livres et est réaliste dans plusieurs applications. La fonction $p_{i,j}(t)$ est la *fonction de transition* de la chaîne de Markov à temps continu.

(ii) Si $p_{i,j}(t) > 0$ pour un $t \geq 0$ et $p_{j,i}(t^*) > 0$ pour un $t^* \geq 0$, on dit que les états i et j *communiquent.* Si tous les états communiquent, la chaîne est dite *irréductible.*

(iii) Dans le contexte de la théorie des files d'attente, les $X(t)$ sont des variables aléatoires non négatives qui prennent des valeurs entières. C'est-à-dire que l'espace des états du processus stochastique $\{X(t), t \geq 0\}$ est l'ensemble $\{0, 1, \ldots\}$. Selon cette hypothèse, on peut écrire que

$$\sum_{j=0}^{\infty} p_{i,j}(t) = 1 \quad \forall i \in \{0, 1, \ldots\}$$

En effet, quel que soit l'état dans lequel se trouve le processus à un instant fixé $\tau \geq 0$, il doit être dans l'un des états possibles à l'instant $\tau + t$, où $t \geq 0$. Notons qu'on a:

$$p_{i,j}(0) = \delta_{i,j} := \begin{cases} 1 \text{ si } i = j \\ 0 \text{ si } i \neq j \end{cases}$$

pour tous les états $i, j \in \{0, 1, \ldots\}$.

Notation. On note $\rho_{i,j}$ la probabilité que la chaîne de Markov à temps continu $\{X(t), t \geq 0\}$, lorsqu'elle quitte l'état courant i, entre dans l'état j, pour $i, j \in \{0, 1, \ldots\}$.

On a, par définition, que $\rho_{i,i} = 0$ pour tous les états i, et

$$\sum_{j=0}^{\infty} \rho_{i,j} = 1 \quad \forall i \in \{0, 1, \ldots\}$$

Définition 9.1.5. *Soit $\{X(t), t \geq 0\}$ une chaîne de Markov à temps continu dont l'espace des états est $\{0, 1, 2, \ldots\}$. Si*

$$\rho_{i,j} = 0 \quad \textit{lorsque } |j - i| > 1 \tag{9.3}$$

le processus est appelé **processus de naissance et de mort**. *De plus, si*

$$\rho_{i,i+1} = 1 \quad \textit{pour tout } i$$

alors $\{X(t), t \geq 0\}$ est un **processus de naissance pur**, *tandis que dans le cas où*

$$\rho_{i,i-1} = 1 \quad \textit{pour tout } i \in \{1, 2, \ldots\}$$

on dit que $\{X(t), t \geq 0\}$ est un **processus de mort pur**.

On déduit de la définition qu'un processus de naissance et de mort est tel que

$$\rho_{0,1} = 1 \quad \text{et} \quad \rho_{i,i+1} + \rho_{i,i-1} = 1 \quad \text{pour } i \in \{1, 2, \ldots\}$$

C'est-à-dire que lorsque le processus est dans l'état $i \geq 1$, le prochain état visité sera nécessairement $i + 1$ ou $i - 1$.

Remarque. L'espace des états d'un processus de naissance et de mort peut être l'ensemble fini $\{0, 1, 2, \ldots, c\}$. On a alors que $\rho_{c,c-1} = 1$.

Dans la théorie des files d'attente, l'état du processus à un instant fixé sera généralement le nombre d'individus dans le système d'attente à cet instant. Lorsque $\{X(t), t \geq 0\}$ passe de l'état i à $i + 1$, on dit qu'une *arrivée* s'est produite, et s'il passe de i à $i - 1$, alors un *départ* a eu lieu. On suppose que, lorsque la chaîne est dans l'état i, le temps A_i requis pour qu'une nouvelle arrivée se produise est une variable aléatoire qui présente une distribution $\text{Exp}(\lambda_i)$, pour $i \in \{0, 1, \ldots\}$. De plus, on suppose que A_i est *indépendant* du temps aléatoire $D_i \sim \text{Exp}(\mu_i)$ écoulé jusqu'au prochain départ, pour $i \in \{1, 2, \ldots\}$.

Proposition 9.1.1. *Le temps total τ_i que le processus de naissance et de mort $\{X(t),\ t \geq 0\}$ passe dans l'état i, lors d'une visite donnée à cet état, est une variable aléatoire qui présente une distribution exponentielle de paramètre*

$$\nu_i = \begin{cases} \lambda_0 & si\ i = 0 \\ \lambda_i + \mu_i & si\ i = 1, 2, \ldots \end{cases}$$

Preuve. Lorsque $i = 0$, on a simplement que $\tau_0 \equiv A_0$, de sorte que $\tau_0 \sim \text{Exp}(\lambda_0)$. Pour $i = 1, 2, \ldots$, on peut écrire que $\tau_i = \min\{A_i, D_i\}$. Le résultat découle alors de la proposition 8.2.1. ■

Remarque. On a aussi (voir l'équation (4.17)) que

$$\rho_{i,i+1} = P[A_i < D_i] = \frac{\lambda_i}{\lambda_i + \mu_i} \quad \text{si } i > 0$$

et

$$\rho_{i,i-1} = P[D_i < A_i] = \frac{\mu_i}{\lambda_i + \mu_i} \quad \text{si } i > 0$$

Définition 9.1.6. *Les paramètres λ_i, pour $i = 0, 1, \ldots$, sont appelés* **taux de naissance** *(ou* **d'arrivée***) du processus de naissance et de mort $\{X(t), t \geq 0\}$, tandis que les paramètres μ_i, pour $i = 1, 2, \ldots$, sont les* **taux de mort** *(ou* **de départ***) du processus.*

Exemple 9.1.1. En plus d'être une chaîne de Markov à temps continu, le processus de Poisson $\{N(t), t \geq 0\}$ est un *processus de comptage* particulier. C'est-à-dire que $N(t)$ symbolise le nombre total d'événements dans l'intervalle $[0, t]$. Puisque seul le nombre d'événements est retenu, et non pas si ces événements étaient des arrivées ou des départs, $\{N(t), t \geq 0\}$ est un exemple de processus de naissance pur. Il s'ensuit que $p_{i,j}(t) = 0$ si $j < i$. De plus, en utilisant le fait que les accroissements du processus de Poisson sont stationnaires, on peut écrire que

$$\begin{aligned} p_{i,j}(t) &\equiv P[N(\tau + t) = j \mid N(\tau) = i] = P[N(t) = j - i] \\ &= P[\text{Poi}(\lambda t) = j - i] = e^{-\lambda t} \frac{(\lambda t)^{j-i}}{(j-i)!} \quad \text{pour } j \geq i \geq 0 \end{aligned}$$

Le temps τ_i que $\{N(t), t \geq 0\}$ passe dans n'importe quel état $i \in \{0, 1, \ldots\}$ présente une distribution exponentielle de paramètre λ. En effet, on a:

$$P[\tau_0 > t] = P[N(t) = 0] = e^{-\lambda t} \quad \text{pour } t \geq 0$$

ce qui implique que $\tau_0 \sim \text{Exp}(\lambda)$. Ensuite, puisque le processus de Poisson possède des accroissements indépendants et stationnaires, on peut alors affirmer que $\tau_i \sim \text{Exp}(\lambda)$, pour $i = 1, 2, \ldots$, également. ♢

En général, il est très difficile de calculer explicitement la fonction de transition $p_{i,j}(t)$. Par conséquent, on doit exprimer les quantités d'intérêt, comme le nombre moyen de clients dans un système de file d'attente donné, en fonction des *probabilités limites* du processus stochastique $\{X(t), t \geq 0\}$.

Définition 9.1.7. *Soit* $\{X(t), t \geq 0\}$ *une chaîne de Markov à temps continu irréductible. La quantité*

$$\pi_j := \lim_{t \to \infty} p_{i,j}(t) \quad \textit{pour tout } j \in \{0, 1, \ldots\}$$

est appelée **probabilité limite** *que le processus soit dans l'état* j *lorsqu'il est* **en équilibre**.

Remarques.
(i) On suppose que les probabilités limites π_j existent et sont indépendantes de l'état initial i.

(ii) Les π_j représentent aussi la proportion du temps que la chaîne de Markov à temps continu passe dans l'état j, sur une longue période.

On peut montrer que les probabilités limites π_j satisfont au système d'équations linéaires suivant:

$$\pi_j \nu_j = \sum_{i \neq j} \pi_i \, \nu_i \rho_{i,j} \quad \forall j \in \{0, 1, \ldots\} \tag{9.4}$$

Pour obtenir les π_j, on peut résoudre le système précédent, sous la condition

$$\sum_{j=0}^{\infty} \pi_j = 1 \tag{9.5}$$

Remarques.
(i) Les diverses équations dans la formule (9.4) sont appelées *équations d'équilibre* du processus stochastique $\{X(t), t \geq 0\}$, parce qu'on peut les interpréter comme suit: le *taux de départ* d'un état j doit être égal au *taux d'arrivée* à j, pour tout j.

(ii) Si $\{X(t), t \geq 0\}$ est un processus de naissance et de mort dont l'espace des états est $\{0, 1, \ldots\}$, les équations d'équilibre sont:

$$\begin{array}{cc} \underline{\text{état } j} & \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j} \\ 0 & \lambda_0\pi_0 = \mu_1\pi_1 \\ 1 & (\lambda_1+\mu_1)\pi_1 = \mu_2\pi_2 + \lambda_0\pi_0 \\ \vdots & \vdots\ \vdots\ \vdots \\ k\ (\geq 1) & (\lambda_k+\mu_k)\pi_k = \mu_{k+1}\pi_{k+1} + \lambda_{k-1}\pi_{k-1} \end{array}$$

Les modèles de base en théorie des files d'attente sont des processus de naissance et de mort particuliers. Pour cette classe de processus, on peut donner la solution générale des équations d'équilibre.

Théorème 9.1.1. *Si* $\{X(t), t \geq 0\}$ *est un processus de naissance et de mort irréductible ayant comme espace des états l'ensemble* $\{0, 1, \ldots\}$*, alors les probabilités limites sont données par*

$$\pi_j = \begin{cases} \dfrac{1}{1+\sum_{k=1}^{\infty} \Pi_k} & \textit{pour } j = 0 \\ \Pi_j \pi_0 & \textit{pour } j = 1, 2, \ldots \end{cases} \tag{9.6}$$

où

$$\Pi_k := \frac{\lambda_0\lambda_1\cdots\lambda_{k-1}}{\mu_1\mu_2\cdots\mu_k} \quad \textit{pour } k \geq 1$$

Remarque. Les probabilités limites existent si et seulement si la somme $\sum_{k=1}^{\infty} \Pi_k$ converge. Dans le cas où l'espace des états de $\{X(t), t \geq 0\}$ est *fini*, la somme en question converge toujours, de sorte que l'existence des probabilités limites est assurée.

Exemple 9.1.2. On suppose que les taux d'arrivée et de départ du processus de naissance et de mort $\{X(t), t \geq 0\}$, dont l'espace des états est $\{0, 1, 2\}$, sont donnés par

$$\lambda_0 = \lambda_1 = \lambda \quad \text{et} \quad \mu_1 = \mu, \quad \mu_2 = 2\mu$$

Écrire les équations d'équilibre du processus et les résoudre pour obtenir les probabilités limites.

Solution. On a:

$$\underline{\text{état } j}\ \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc} 0 & \lambda\pi_0 = \mu\pi_1 \\ 1 & (\lambda+\mu)\pi_1 = 2\mu\pi_2 + \lambda\pi_0 \\ 2 & 2\mu\pi_2 = \lambda\pi_1 \end{array}$$

Puisque ce système d'équations est simple, on peut le résoudre facilement. On déduit de l'équation de l'état 0 que

$$\pi_1 = \frac{\lambda}{\mu}\pi_0$$

De façon similaire, l'équation de l'état 2 implique que

$$\pi_2 = \frac{\lambda}{2\mu}\pi_1 = \left(\frac{\lambda}{2\mu}\right)\left(\frac{\lambda}{\mu}\right)\pi_0$$

Il s'ensuit que

$$\pi_0 + \frac{\lambda}{\mu}\pi_0 + \left(\frac{\lambda}{2\mu}\right)\left(\frac{\lambda}{\mu}\right)\pi_0 = 1$$

C'est-à-dire que

$$\pi_0 = \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{2\mu}\right)\left(\frac{\lambda}{\mu}\right)\right]^{-1}$$

de sorte que

$$\pi_1 = \left(\frac{\lambda}{\mu}\right)\left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{2\mu}\right)\left(\frac{\lambda}{\mu}\right)\right]^{-1}$$

et

$$\pi_2 = \left(\frac{\lambda}{2\mu}\right)\left(\frac{\lambda}{\mu}\right)\left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{2\mu}\right)\left(\frac{\lambda}{\mu}\right)\right]^{-1}$$

Remarques.

(i) On peut vérifier que l'équation de l'état 1, dont on ne s'est pas servi pour résoudre le système d'équations linéaires, est aussi satisfaite par la solution obtenue ci-dessus.

(ii) Puisque $\{X(t), t \geq 0\}$ est un processus de naissance et de mort particulier, on peut aussi faire appel au théorème 9.1.1 pour trouver les probabilités limites. On a:

$$\Pi_1 := \frac{\lambda}{\mu} \quad \text{et} \quad \Pi_2 := \frac{\lambda \times \lambda}{\mu \times 2\mu}$$

d'où l'on retrouve les formules pour π_0, π_1 et π_2. ♢

9.2 Systèmes de files d'attente avec un seul serveur

Soit $X(t)$ le nombre de *clients* dans un *système de file d'attente* à l'instant t. Si l'on suppose que les temps A_n entre les arrivées des clients successifs et les temps de service S_n des clients sont des variables aléatoires indépendantes qui présentent une distribution exponentielle, alors le processus $\{X(t), t \geq 0\}$ est une chaîne de Markov à temps continu. De plus, dans la plupart des cas, on suppose aussi que les clients arrivent un à la fois et sont servis un à la fois. Il s'ensuit que $\{X(t), t \geq 0\}$ est un processus de naissance et de mort. Les arrivées des clients dans le système constituent un processus de Poisson. On peut montrer que les départs du système *en équilibre* constituent également un processus de Poisson. Un tel système de file d'attente est noté $M/M/s$, où s est le nombre de *serveurs* dans le système. Dans cette section, s est égal à 1.

Remarques.
(i) Nous avons utilisé le mot *clients* ci-dessus. Cependant, les clients dans un système de file d'attente peuvent en fait être, par exemple, des machines dans un atelier de réparation, ou des travaux dans un système informatique, ou encore des avions qui décollent d'un aéroport ou y atterrissent, etc.

(ii) Pour plus de précision, il faudrait spécifier que les variables aléatoires S_n sont indépendantes des A_n. De plus, les S_n sont des variables aléatoires identiquement distribuées, de même que les A_n.

(iii) La notation M pour le processus d'arrivée (et le processus de départ) est utilisée parce que le processus de Poisson est *markovien*.

On s'intéresse au *nombre moyen* de clients et au *temps moyen* qu'un client quelconque passe dans le système de file d'attente, lorsque ce système est en *équilibre* ou en *régime stationnaire*.

Notations. On note respectivement $\bar{N}$, $\bar{N}_Q$ et $\bar{N}_S$ le nombre moyen (total) de clients dans le système en équilibre, le nombre moyen de clients qui font la queue et le nombre moyen de clients en train d'être servis. De plus, $\bar{T}$ est le temps moyen (total) qu'un client quelconque passe dans le système, $\bar{Q}$ est son temps d'attente moyen et $\bar{S}$ est son temps moyen de service.

On a que $\bar{N} = \bar{N}_Q + \bar{N}_S$ et $\bar{T} = \bar{Q} + \bar{S}$. Comme nous l'avons mentionné dans la section précédente, on peut exprimer les diverses quantités d'intérêt en fonction des probabilités limites π_n du processus stochastique $\{X(t), t \geq 0\}$.

Définition 9.2.1. *Soit $N(t)$, pour $t \geq 0$, le nombre d'arrivées dans le système dans l'intervalle $[0, t]$. La quantité*

$$\lambda_a := \lim_{t\to\infty} \frac{N(t)}{t} \tag{9.7}$$

est appelée **taux moyen d'arrivée** *des clients dans le système.*

Remarques.
(i) On peut montrer que

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{E[A_n]}$$

Dans le cas présent, on suppose que $A_n \sim \text{Exp}(\lambda)$, pour $n = 0, 1, \ldots$, de sorte que le processus stochastique $\{N(t), t \geq 0\}$ est un processus de Poisson de taux $\lambda > 0$. Il s'ensuit que $\lambda_a = \lambda$.

(ii) Lorsque la capacité du système est infinie, tous les clients qui arrivent peuvent entrer dans le système. Cependant, en pratique, la capacité de n'importe quel système est finie. Par conséquent, on considère aussi le *taux moyen d'entrée* des clients dans le système, lequel est noté λ_e. Dans le cas où la capacité du système est égale à une constante c ($< \infty$), on a que $\lambda_e = \lambda(1 - \pi_c)$, car $(1 - \pi_c)$ est la probabilité (limite) qu'un client qui arrive puisse entrer dans le système. Notons que, en fait, même si la capacité du système est supposée infinie, certains clients qui arrivent peuvent décider de ne pas entrer dans le système, par exemple s'ils trouvent que la longueur de la file d'attente est trop grande. Ainsi, en général, λ_e est inférieur ou égal à λ.

(iii) Soit $D(t)$ le nombre de départs du système de file d'attente dans l'intervalle $[0, t]$. On suppose que

$$\lambda_d := \lim_{t\to\infty} \frac{D(t)}{t} = \lambda_e$$

Pour analyser un système de file d'attente donné, on commence souvent par calculer ses probabilités limites π_n. Ensuite, on essaie d'obtenir les quantités $\bar{N}$, $\bar{N}_Q$, et ainsi de suite, en fonction des π_n. De plus, on peut utiliser une *équation de coût* pour établir une relation entre $\bar{N}$ et $\bar{T}$. En effet, si on suppose qu'un client quelconque paie 1 $ par unité de temps qu'il passe dans le système (à faire la queue ou à être servi), alors on peut montrer que

$$\bar{N} = \lambda_e \cdot \bar{T} \tag{9.8}$$

Cette équation est connue sous le nom de *formule de Little* (ou loi de Little). Elle est valable si l'on suppose que λ_e et $\bar{T}$ existent et sont finis. De plus, on a:

$$\bar{N} = \lim_{t\to\infty} \frac{1}{t} \int_0^t X(s)ds$$

et, si T_k symbolise le temps passé dans le système par le k^{e} client,

$$\bar{T} = \lim_{t\to\infty} \frac{\sum_{k=1}^{N(t)} T_k}{N(t)}$$

Remarques.
(i) La formule de Little est valable pour des systèmes très généraux, en particulier pour les systèmes $M/M/s$, de capacité finie ou infinie, qui sont étudiés dans ce livre.
(ii) Lorsque t est assez grand pour que le processus soit en régime stationnaire, on peut écrire que

$$\bar{N} = E[X(t)]$$

En plus de la formule (9.8), on a aussi:

$$\bar{N}_S = \lambda_e \cdot \bar{S} \tag{9.9}$$

Il s'ensuit, en utilisant le fait que $\bar{N} = \bar{N}_Q + \bar{N}_S$ et $\bar{T} = \bar{Q} + \bar{S}$, que

$$\bar{N}_Q = \lambda_e \cdot \bar{Q}$$

Dans le cas du modèle $M/M/s$, $\lambda_e = \lambda$ et les temps de service S_n sont supposés être des variables aléatoires i.i.d. présentant une distribution exponentielle de paramètre μ. De là, on déduit que $\bar{S} = E[S_n] = 1/\mu$. L'équation (9.9) implique alors que $\bar{N}_S = \lambda/\mu$.

9.2.1 Modèle $M/M/1$

Le système de file d'attente de base est le modèle $M/M/1$. Dans ce modèle, on suppose que les arrivées successives des clients constituent un processus de Poisson de taux λ et que les temps de service S_n sont des variables aléatoires $\text{Exp}(\mu)$ indépendantes. De plus, les S_n sont indépendants des temps entre les arrivées des clients. Finalement, la capacité du système est infinie, et on tient pour acquis que tous les clients qui arrivent décident d'entrer dans le système, peu importe l'état de ce système à leur arrivée.

Le processus stochastique $\{X(t), t \geq 0\}$, où $X(t)$ représente le nombre de clients dans le système à l'instant $t \geq 0$, est un processus de naissance et de

mort irréductible. En effet, puisque les taux d'arrivée $\lambda_n \equiv \lambda$ et les taux de départ $\mu_n \equiv \mu$ sont strictement positifs pour n'importe quelle valeur de n, tous les états communiquent. On trouve que les équations d'équilibre pour la file d'attente $M/M/1$ sont (voir la page 469):

$$\underline{\text{état } j} \quad \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc} 0 & \lambda\pi_0 = \mu\pi_1 \\ n\ (\geq 1) & (\lambda+\mu)\pi_n = \lambda\pi_{n-1} + \mu\pi_{n+1} \end{array}$$

On peut résoudre le système d'équations linéaires précédent, sous la condition $\sum_{n=0}^{\infty} \pi_n = 1$, pour obtenir les probabilités limites. Cependant, le théorème 9.1.1 donne la solution presque immédiatement. On calcule

$$\Pi_k = \underbrace{\frac{\lambda\lambda\cdots\lambda}{\mu\mu\cdots\mu}}_{k \text{ fois}} = \left(\frac{\lambda}{\mu}\right)^k \quad \text{pour } k = 1, 2, \ldots \tag{9.10}$$

Il s'ensuit que

$$S^* := \sum_{k=1}^{\infty} \Pi_k = \frac{\lambda/\mu}{1-(\lambda/\mu)} < \infty \quad \text{si et seulement si} \quad \rho := \frac{\lambda}{\mu} < 1$$

Remarques.
(i) La quantité ρ est appelée *intensité du trafic* ou *taux d'utilisation* du système. Puisque $1/\mu$ est le temps moyen de service d'un client quelconque et λ est le taux moyen d'arrivée des clients, la condition $\rho < 1$ signifie que les clients ne doivent pas arriver plus rapidement que le taux auquel ils sont servis ou, de façon équivalente, plus rapidement que le temps moyen que cela prend pour servir un client, si l'on veut que le système atteigne un régime *stationnaire* (ou *permanent*). Lorsque $\rho \geq 1$, on peut affirmer que la longueur de la file d'attente augmentera indéfiniment.

(ii) Au chapitre 2, nous avons utilisé des diagrammes de Venn pour représenter des espaces échantillons et des événements. Dans la théorie des files d'attente, on dessine un *diagramme de transitions* pour décrire un système donné. Les états possibles du système sont représentés par des cercles. Pour indiquer qu'une transition de l'état i à l'état j est possible, on trace une flèche allant du cercle correspondant à l'état i à celui représentant j. On écrit aussi au-dessus (ou au-dessous) de chaque flèche le taux de la transition en question (voir la figure 9.1).

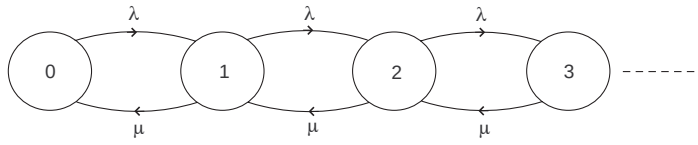

Fig. 9.1. Diagramme de transitions pour le modèle $M/M/1$

Une fois que le diagramme de transitions approprié a été dessiné, il est facile d'écrire les équations d'équilibre du système.

Ensuite, on déduit du théorème 9.1.1 que, si $\rho < 1$,

$$\pi_0 = \frac{1}{1+S^*} = \left(\frac{1}{1-(\lambda/\mu)}\right)^{-1} = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

et

$$\pi_j = \Pi_j \pi_0 = \left(\frac{\lambda}{\mu}\right)^j \left(1 - \frac{\lambda}{\mu}\right) \quad \text{pour } j = 1, 2, \ldots$$

C'est-à-dire que

$$\pi_k = \rho^k (1-\rho) \quad \forall k \geq 0 \tag{9.11}$$

En utilisant les probabilités limites et la formule (1.7) avec $a = 1-\rho$ et $r = \rho$, on peut écrire que

$$\bar{N} := \sum_{k=0}^{\infty} k\pi_k = \sum_{k=0}^{\infty} k\rho^k (1-\rho) = \frac{(1-\rho)\rho}{(1-\rho)^2} = \frac{\rho}{1-\rho} \tag{9.12}$$

Remarques.
(i) Notons que $\lim_{\rho \uparrow 1} \bar{N} = \infty$, ce qui correspond au fait que la longueur de la file d'attente augmente indéfiniment si $\rho = 1$ (et, *a fortiori*, si $\rho > 1$).
(ii) Si on note N le nombre (aléatoire) de clients dans le système en équilibre, de sorte que $\bar{N} = E[N]$, on peut écrire que $N_1 := N + 1$ présente une distribution géométrique de paramètre $p := 1 - \rho$.

Maintenant, on déduit de la formule de Little de l'équation (9.8) que

$$\bar{T} = \frac{\bar{N}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu - \lambda} \tag{9.13}$$

Puisque $\bar{S} = 1/\mu$ et $\bar{N}_S = \lambda/\mu = \rho$, comme nous l'avons déjà mentionné ci-dessus, il s'ensuit que

$$\bar{Q} = \bar{T} - \bar{S} = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu - \lambda)} \tag{9.14}$$

et

$$\bar{N}_Q = \bar{N} - \bar{N}_S = \frac{\rho}{1 - \rho} - \rho = \frac{\rho^2}{1 - \rho} \tag{9.15}$$

Remarque. Soit N_S la variable aléatoire qui représente le nombre de clients en train d'être servis lorsque le système est en régime stationnaire. Parce qu'il n'y a qu'un seul serveur, N_S présente une distribution de Bernoulli de paramètre $p_0 := 1 - \pi_0$, ce qui implique que

$$\bar{N}_S = E[N_S] = p_0 = 1 - \pi_0 = 1 - (1 - \rho) = \rho$$

comme nous l'avons rappelé ci-dessus.

Nous avons donné dans ce qui précède les distributions exactes des variables aléatoires S, N et N_S. On peut aussi trouver les distributions des variables T, Q et N_Q, où T est le temps total qu'un client quelconque passe dans le système (en équilibre), Q est le temps d'attente et N_Q est le nombre de clients qui attendent d'être servis. On a déjà déterminé que $E[T] = \bar{T} = 1/(\mu - \lambda)$. En fait, T présente une distribution exponentielle, et la quantité $\mu - \lambda$ est son paramètre.

Proposition 9.2.1. *Le temps total T qu'un client quelconque passe dans une file d'attente $M/M/1$ en équilibre est une variable aléatoire qui présente une distribution exponentielle de paramètre $\mu - \lambda$.*

Preuve. Pour démontrer le résultat, on conditionne sur le nombre K de clients déjà dans le système à l'arrivée du client d'intérêt. On peut écrire que

$$P[T \le t] = \sum_{k=0}^{\infty} P[T \le t \mid K = k] P[K = k]$$

Maintenant, par la *propriété de non-vieillissement* de la distribution exponentielle, si $K = k$, alors la variable aléatoire T est la somme de $k + 1$ variables aléatoires indépendantes qui présentent toutes une distribution $\mathrm{Exp}(\mu)$. En effet, le temps de service du client en train d'être servi (si $k > 0$), à partir du moment où le client d'intérêt entre dans le système, présente aussi une distribution exponentielle de paramètre μ. En utilisant l'équation (4.27), on peut écrire que

$$T \mid \{K = k\} \sim \mathrm{G}(k + 1, \mu)$$

Ensuite, on peut montrer que lorsque le processus d'arrivée est un processus de Poisson, la probabilité $P[K = k]$ qu'un client quelconque trouve k clients dans le système en équilibre à son arrivée est égale à la probabilité limite π_k qu'il y ait k clients dans le système (en équilibre). Il s'ensuit que

$$P[K = k] = \pi_k = \rho^k(1 - \rho) \quad \text{pour } k = 0, 1, \ldots$$

De là, on a:

$$\begin{aligned} P[T \le t] &= \sum_{k=0}^{\infty} \left[\int_0^t \mu e^{-\mu\tau} \frac{(\mu\tau)^k}{k!} d\tau \right] \rho^k (1 - \rho) \\ &\overset{\rho=\lambda/\mu}{=} (\mu - \lambda) \sum_{k=0}^{\infty} \int_0^t e^{-\mu\tau} \frac{(\lambda\tau)^k}{k!} d\tau = (\mu - \lambda) \int_0^t e^{-\mu\tau} \sum_{k=0}^{\infty} \frac{(\lambda\tau)^k}{k!} d\tau \\ &= (\mu - \lambda) \int_0^t e^{-\mu\tau} e^{\lambda\tau} d\tau = 1 - e^{-(\mu-\lambda)t} \end{aligned}$$

ce qui implique que

$$f_T(t) = \frac{d}{dt} P[T \le t] = (\mu - \lambda) e^{-(\mu-\lambda)t} \quad \text{pour } t \ge 0$$

■

Le temps d'attente, Q, d'un client quelconque qui entre dans le système est une variable aléatoire de type *mixte*. En utilisant le fait que $P[K = k] = \pi_k$, où K est défini ci-dessus, on a:

$$P[Q = 0] = P[K = 0] = \pi_0 = 1 - \rho$$

Dans le cas où $K = k > 0$, on peut écrire que $Q \sim \mathrm{G}(k, \mu)$. Alors, en conditionnant sur les valeurs possibles de la variable aléatoire K, on obtient que

$$P[Q \le t] = 1 - \left(\frac{\lambda}{\mu} \right) e^{-(\mu-\lambda)t} \quad \text{pour } t \ge 0$$

Remarque. On trouve que $Q \mid \{Q > 0\} \sim \mathrm{Exp}(\mu - \lambda)$. C'est-à-dire que $R := Q \mid \{Q > 0\}$ et le temps total T qu'un client quelconque passe dans le système sont des variables aléatoires identiquement distribuées. Notons que puisqu'une variable aléatoire exponentielle est continue, on peut la définir indifféremment dans l'intervalle $[0, \infty)$ ou l'intervalle $(0, \infty)$.

Finalement, le nombre N_Q de clients faisant la queue lorsque le système est en régime stationnaire peut être exprimé comme suit:

$$N_Q = \begin{cases} 0 & \text{si } N = 0 \\ N-1 & \text{si } N = 1, 2, \ldots \end{cases}$$

De là, on peut écrire que

$$P[N_Q = 0] = P[N = 0] + P[N = 1] = \pi_0 + \pi_1 = (1+\rho)(1-\rho)$$

et

$$P[N_Q = k] = P[N = k+1] = \pi_{k+1} = \rho^{k+1}(1-\rho) \quad \text{si } k = 1, 2, \ldots$$

Remarques.
(i) Puisqu'on a obtenu les distributions de toutes les variables aléatoires d'intérêt, on pourrait calculer leurs variances respectives.

(ii) Les variables aléatoires Q et S sont indépendantes. Cependant, N_Q et N_S ne sont *pas* indépendantes. En effet, on peut écrire que

$$N_S = 0 \quad \Longrightarrow \quad N_Q = 0$$

(parce que si personne n'est en train d'être servi, alors personne ne fait la queue non plus).

Exemple 9.2.1. On suppose qu'à un instant fixé $t_0 > 0$, le nombre $X(t_0)$ de clients dans une file d'attente $M/M/1$ en régime stationnaire est inférieur ou égal à trois. Calculer l'espérance mathématique de la variable aléatoire $X(t_0)$, de même que sa variance, si $\lambda = \mu/2$.

Solution. Puisque $\rho = 1/2$, les probabilités limites du système sont:

$$\pi_k = \rho^k(1-\rho) = (1/2)^{k+1} \quad \text{pour } k = 0, 1, 2, \ldots$$

Il s'ensuit que

$$P[X(t_0) \leq 3] = \sum_{k=0}^{3} (1/2)^{k+1} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16}$$

Par conséquent, sous la condition $X(t_0) \leq 3$, on a:

$$\pi_0 = \frac{1/2}{15/16} = \frac{8}{15}, \quad \pi_1 = \frac{4}{15}, \quad \pi_2 = \frac{2}{15} \quad \text{et} \quad \pi_3 = \frac{1}{15}$$

Il est alors facile d'obtenir la moyenne et la variance de $X(t_0)$. On calcule

$$E[X(t_0)] = 0 + 1 \times \frac{4}{15} + 2 \times \frac{2}{15} + 3 \times \frac{1}{15} = \frac{11}{15}$$

et

$$E[X^2(t_0)] = 0 + 1^2 \times \frac{4}{15} + 2^2 \times \frac{2}{15} + 3^2 \times \frac{1}{15} = \frac{21}{15}$$

de sorte que

$$\mathrm{VAR}[X(t_0)] = \frac{21}{15} - \left(\frac{11}{15}\right)^2 = \frac{194}{225}$$

Remarque. La condition $X(t_0) \leq 3$ ne signifie pas que la capacité du système est de $c = 3$ clients. Lorsque $c = 3$, la variable aléatoire $X(t)$ doit nécessairement être inférieure ou égale à trois pour toutes les valeurs de t, tandis que dans l'exemple ci-dessus, le nombre de clients dans le système *à un instant fixé* était inférieur à quatre. Le cas où la capacité du système est finie est le sujet de la prochaine sous-section. ♢

Exemple 9.2.2. Souvent, un modèle de file d'attente particulier est une modification plus ou moins simple du modèle de base $M/M/1$. Par exemple, supposons que le serveur, dans un système de file d'attente qui, sans cette modification, serait le modèle $M/M/1$, attend toujours jusqu'à ce qu'il y ait (au moins) deux clients dans le système avant de commencer à les servir, exactement deux à la fois, à un taux exponentiel μ (c'est-à-dire que le temps de service présente une distribution exponentielle de paramètre μ). Alors le processus stochastique $\{X(t), t \geq 0\}$, où $X(t)$ symbolise le nombre de clients dans le système à l'instant $t \geq 0$, est encore une chaîne de Markov à temps continu. Cependant, il n'est plus un processus de naissance et de mort. Pour obtenir les probabilités limites du processus, il faut résoudre les équations d'équilibre appropriées, sous la condition $\sum_{k=0}^{\infty} \pi_k = 1$. Ces équations d'équilibre sont (voir la figure 9.2):

état j	taux de départ de j $=$ taux d'arrivée à j
0	$\lambda\pi_0 \overset{(0)}{=} \mu\pi_2$
1	$\lambda\pi_1 \overset{(1)}{=} \lambda\pi_0 + \mu\pi_3$
2	$(\lambda+\mu)\pi_2 \overset{(2)}{=} \lambda\pi_1 + \mu\pi_4$
k (≥ 3)	$(\lambda+\mu)\pi_k = \lambda\pi_{k-1} + \mu\pi_{k+2}$

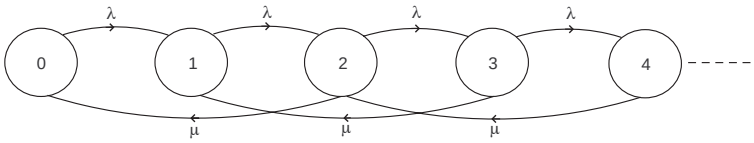

Fig. 9.2. Diagramme de transitions pour le modèle de file d'attente de l'exemple 9.2.2

Une solution de l'équation pour $k \geq 3$ peut être obtenue en supposant que $\pi_k = a^{k-2}\pi_2$, pour $k = (2), 3, 4, \ldots$, où a est une constante telle que $0 < a < 1$. L'équation en question devient:

$$(\lambda + \mu)a^{k-2}\pi_2 = \lambda a^{k-3}\pi_2 + \mu a^k \pi_2$$

Étant donné que la probabilité π_2 ne peut pas être nulle, on peut écrire que

$$(\lambda + \mu)a = \lambda + \mu a^3$$

C'est-à-dire qu'il faut résoudre une équation polynomiale du troisième degré. On trouve que $a = 1$ est une racine évidente. Il s'ensuit que

$$(\lambda + \mu)a = \lambda + \mu a^3 \quad \Longleftrightarrow \quad (a-1)(\mu a^2 + \mu a - \lambda) = 0$$

De là, les deux autres racines sont:

$$a = -\frac{1}{2} \pm \frac{\sqrt{\mu^2 + 4\mu\lambda}}{2\mu}$$

Puisque $a > 0$, on déduit que

$$a = -\frac{1}{2} + \frac{\sqrt{\mu^2 + 4\mu\lambda}}{2\mu} = \frac{1}{2}\left(\sqrt{1 + 4\rho} - 1\right)$$

Remarques.
(i) La solution $a = 1$ doit être éliminée, car elle impliquerait que $\pi_k = \pi_2$, pour $k = 2, 3, \ldots$, de sorte que la condition $\sum_{k=0}^{\infty} \pi_k = 1$ ne pourrait pas être satisfaite.
(ii) On doit aussi avoir:

$$\frac{1}{2}\left(\sqrt{1 + 4\rho} - 1\right) < 1 \quad \Longleftrightarrow \quad \sqrt{1 + 4\rho} < 3$$

De là, on déduit que les probabilités limites existent (si et) seulement si $\rho < 2$ ou, de façon équivalente, si $\lambda < 2\mu$. C'est-à-dire que le taux d'arrivée des clients ne doit pas excéder leur taux de service. Il s'agit de la même condition de l'existence des probabilités limites que celle qu'on a dans le cas du modèle $M/M/2$, comme nous le démontrerons dans la prochaine section.

Pour compléter cet exemple, on utilise les équations des états 0 et 1 ci-dessus pour exprimer π_0 et π_1 en fonction de π_2. L'équation (0) implique directement que $\pi_0 = (\mu/\lambda)\pi_2$, tandis que (0) et (1) ensemble impliquent que

$$\lambda\pi_1 = \mu\pi_2 + \mu\pi_3 = \mu\pi_2 + \mu a\pi_2 \quad \Longrightarrow \quad \pi_1 = \frac{\mu}{\lambda}(a+1)\pi_2$$

Alors la condition $\sum_{k=0}^{\infty} \pi_k = 1$ permet d'obtenir une expression explicite de π_2 (de laquelle on déduit la valeur de π_k, pour $k = 0, 1, 3, 4, \ldots$). On a:

$$\begin{aligned} 1 &= \sum_{k=0}^{\infty} \pi_k = \frac{\mu}{\lambda}\pi_2 + \frac{\mu}{\lambda}(a+1)\pi_2 + \sum_{k=2}^{\infty} a^{k-2}\pi_2 \\ &= \pi_2 \left[\frac{\mu}{\lambda}(a+2) + \sum_{k=0}^{\infty} a^k\right] = \pi_2 \left[\frac{\mu}{\lambda}(a+2) + \frac{1}{1-a}\right] \end{aligned}$$

Donc, on peut écrire que

$$\pi_2 = \left[\frac{\mu}{\lambda}(a+2) + \frac{1}{1-a}\right]^{-1}$$

Observons qu'on ne s'est pas servi de l'équation de l'état 2 pour déterminer les probabilités limites π_k. En fait, il y a toujours une équation redondante dans le système d'équations linéaires. On peut maintenant vérifier que la solution obtenue satisfait aussi à l'équation (2) ci-dessus. On a:

$$(\lambda+\mu)\pi_2 = \lambda\pi_1 + \mu\pi_4 \quad \Longleftrightarrow \quad (\lambda+\mu)\pi_2 = \lambda\frac{\mu}{\lambda}(a+1)\pi_2 + \mu a^2\pi_2$$

C'est-à-dire qu'on doit avoir:

$$\mu a^2 + \mu a - \lambda = 0$$

Mais il s'agit exactement de l'équation quadratique que vérifie la constante a (voir ci-dessus).

Remarques.
(i) On a obtenu *une* solution des équations d'équilibre, sous la condition de normalisation (9.5). En fait, on peut montrer qu'il y a une *unique* solution de ce système d'équations linéaires qui satisfait à la condition (9.5). Par conséquent, on peut affirmer qu'on a trouvé *la* solution du problème.

(ii) Si le serveur est *capable* de servir deux clients à la fois (également au taux μ), mais commence à travailler dès qu'il y a un client dans le système, alors la solution est légèrement différente (voir la référence [26]).

(iii) Si l'on suppose plutôt que les clients arrivent toujours deux à la fois, mais ne sont servis qu'un à la fois, alors les équations d'équilibre deviennent:

$$\underline{\text{état } j} \quad \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc} 0 & \lambda\pi_0 = \mu\pi_1 \\ 1 & (\lambda+\mu)\pi_1 = \mu\pi_2 \\ k\ (\geq 2) & (\lambda+\mu)\pi_k = \lambda\pi_{k-2} + \mu\pi_{k+1} \end{array}$$

Dans un tel cas, on pourrait déterminer au hasard les positions respectives dans la file d'attente des deux clients qui sont arrivés ensemble.

(iv) Finalement, si les clients arrivent toujours deux à la fois et sont aussi toujours servis deux à la fois, alors les probabilités limites π_n^* de la chaîne de Markov à temps continu correspondante peuvent être exprimées en fonction des probabilités limites π_n du modèle $M/M/1$ comme suit:

$$\pi_n^* = \pi_{n/2} = \rho^{n/2}(1-\rho) \quad \text{pour } n = 0, 2, 4, \ldots$$

♢

9.2.2 Modèle $M/M/1$ à capacité finie

Comme nous l'avons déjà mentionné, en pratique, la capacité de n'importe quel système de file d'attente est limitée. Soit c l'entier représentant cette capacité. Supposons qu'on a calculé les probabilités limites d'un système de file d'attente donné ayant une capacité finie et qu'on a trouvé que la valeur de π_c est très petite. Alors, supposer que c est en fait l'infini est une approximation simplificatrice valable. Cependant, si la probabilité que le système soit *saturé* est loin d'être négligeable, alors on devrait utiliser un espace des états fini.

Supposons qu'un certain système de file d'attente peut être décrit adéquatement par un modèle $M/M/1$ ayant $c+1$ états possibles: $0, 1, \ldots, c$. Ce modèle

est souvent noté $M/M/1/c$. Les équations d'équilibre du système sont alors les suivantes:

$$\begin{array}{cc} \underline{\text{état } j} & \underline{\text{taux de départ de } j = \text{taux d'arrivée à } j} \\ 0 & \lambda\pi_0 = \mu\pi_1 \\ k = 1, \ldots, c-1 & (\lambda+\mu)\pi_k = \lambda\pi_{k-1} + \mu\pi_{k+1} \\ c & \mu\pi_c = \lambda\pi_{c-1} \end{array}$$

Notons que les équations d'équilibre pour les états $j = 0, 1, \ldots, c-1$ sont identiques aux équations correspondantes dans la file d'attente $M/M/1/\infty$. Lorsque le système a atteint sa capacité maximale, c'est-à-dire c clients, le prochain état visité sera nécessairement $c-1$, à un taux exponentiel μ. De plus, la seule façon dont le système peut entrer dans l'état c est à partir de l'état $c-1$, lorsqu'un nouveau client arrive.

Soit encore une fois $X(t)$ le nombre de clients dans le système à l'instant $t \geq 0$. Le processus stochastique $\{X(t), t \geq 0\}$ est un processus de naissance et de mort irréductible, comme précédemment. Par conséquent, plutôt que de résoudre le système d'équations linéaires ci-dessus, sous la condition $\sum_{j=0}^{c} \pi_j = 1$ (voir l'équation (9.5)), on peut faire appel au théorème 9.1.1. On a encore:

$$\Pi_k = \left(\frac{\lambda}{\mu}\right)^k = \rho^k \quad \text{pour } k = 1, 2, \ldots, c$$

de sorte que

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{c} \Pi_k} = \frac{1}{1 + \sum_{k=1}^{c} \rho^k}$$

et

$$\pi_j = \Pi_j \pi_0 = \frac{\rho^j}{1 + \sum_{k=1}^{c} \rho^k} = \frac{\rho^j}{\sum_{k=0}^{c} \rho^k} \quad \text{pour } j = 1, 2, \ldots, c$$

Parce que l'espace des états est fini, les probabilités limites existent pour n'importe quelles valeurs (positives) des paramètres λ et μ. Dans le cas particulier où $\rho = 1$, la solution est simplement:

$$\pi_j = \frac{1}{c+1} \quad \text{pour } j = 0, 1, \ldots, c \tag{9.16}$$

C'est-à-dire que lorsque le système est en équilibre, les $c+1$ états possibles de la chaîne de Markov sont équiprobables.

Lorsque $\rho \neq 1$, on calcule

$$\sum_{k=0}^{c} \rho^k = \frac{1-\rho^{c+1}}{1-\rho}$$

De là, on peut écrire que

$$\pi_j \overset{\rho \neq 1}{=} \frac{\rho^j(1-\rho)}{1-\rho^{c+1}} \quad \text{pour } j = 0, 1, \ldots, c \tag{9.17}$$

Remarque. La probabilité que le système soit saturé est donnée par

$$\pi_c = \frac{\rho^c(1-\rho)}{1-\rho^{c+1}}$$

En prenant la limite lorsque ρ tend vers l'infini, on obtient que

$$\lim_{\rho \to \infty} \pi_c = \lim_{\rho \to \infty} \frac{\rho^c(1-\rho)}{1-\rho^{c+1}} = \lim_{\rho \to \infty} \frac{\rho^{-1}-1}{\rho^{-(c+1)}-1} = 1$$

de sorte que $\pi_j = 0$, pour $j = 0, 1, \ldots, c-1$, comme on aurait pu l'anticiper. Inversement, on a:

$$\lim_{\rho \downarrow 0} \pi_0 = \lim_{\rho \downarrow 0} \frac{1-\rho}{1-\rho^{c+1}} = 1$$

et $\pi_j = 0$, pour $j = 1, 2, \ldots, c$. Finalement, si $\rho < 1$ et c tend vers l'infini, on retrouve la formule

$$\pi_j = \lim_{c \to \infty} \frac{\rho^j(1-\rho)}{1-\rho^{c+1}} = \rho^j(1-\rho) \quad \text{pour } j = 0, 1, \ldots$$

obtenue dans le cas du modèle $M/M/1/\infty$.

En utilisant l'équation (9.16), on trouve facilement que

$$\bar{N} = \frac{c}{2} \quad \text{si } \rho = 1$$

Dans le cas général où $\rho \neq 1$, on peut montrer que

$$\bar{N} = \frac{\rho}{1-\rho} - \frac{(c+1)\rho^{c+1}}{1-\rho^{c+1}}$$

En fait, lorsque la capacité c du système est faible, il est facile de calculer la valeur de $\bar{N}$ à partir de la formule

$$\bar{N} \equiv E[N] := \sum_{k=0}^{c} k\pi_k$$

De même, après avoir calculé les π_k, il n'est pas difficile d'obtenir la variance de la variable aléatoire N.

Ensuite, puisque N_S est égal à 1 si le système en équilibre est dans n'importe quel état $k \in \{1, 2, \ldots, c\}$ (et à 0 si le système est vide), l'expression de la valeur de $\bar{N}_S$ est la même qu'auparavant, soit:

$$\bar{N}_S = 1 - \pi_0$$

ce qui implique que

$$\bar{N}_Q = \bar{N} - 1 + \pi_0$$

Cependant, la probabilité limite π_0 est différente de la probabilité limite correspondante dans le modèle $M/M/1/\infty$.

Finalement, si on ne tient compte que des clients qui entrent vraiment dans le système (en équilibre), on peut écrire que leur taux moyen d'entrée est

$$\lambda_e = \lambda(1 - \pi_c)$$

On déduit alors de la formule de Little (voir l'équation (9.8)) que

$$\bar{T} = \frac{\bar{N}}{\lambda(1 - \pi_c)}$$

de sorte que

$$\bar{Q} = \frac{\bar{N}}{\lambda(1 - \pi_c)} - \frac{1}{\mu}$$

car $\bar{S} \equiv E[S] = 1/\mu$, comme auparavant.

Exemple 9.2.3. Considérons le système de file d'attente $M/M/1/2$. C'est-à-dire que la capacité du système est $c = 2$. Supposons que $\lambda = \mu$.
(a) Quelle est la variance du nombre de clients dans le système en régime stationnaire?

(b) Quel est le nombre moyen d'arrivées dans le système (en régime stationnaire) pendant le temps de service d'un client donné?

Solution. (a) On déduit de l'équation (9.16) que $\pi_0 = \pi_1 = \pi_2 = 1/3$. Il s'ensuit que

$$E[N] = \frac{1}{3}(0+1+2) = 1 \quad \text{et} \quad E[N^2] = \frac{1}{3}(0+1+4) = \frac{5}{3}$$

de sorte que

$$\text{VAR}[N] = \frac{5}{3} - 1^2 = \frac{2}{3}$$

(b) Soit $t_0 > 0$ l'instant auquel le client en question commence à être servi. Alors $X(t_0)$ est égal à 1 ou 2. Parce que $\pi_k \equiv 1/3$, on peut affirmer que

$$P[X(t_0) = 1 \mid X(t_0) \in \{1,2\}] = P[X(t_0) = 2 \mid X(t_0) \in \{1,2\}] = \frac{1}{2}$$

Ensuite, soit K le nombre de clients qui entrent dans le système pendant que le client d'intérêt est en train d'être servi. Puisque $c = 2$, les valeurs possibles de la variable aléatoire K sont 0 et 1. C'est-à-dire que K est une variable aléatoire de Bernoulli. On a, sous la condition que $X(t_0) \in \{1,2\}$:

$$\begin{aligned} P[K=0] &= \frac{1}{2}\left\{P[K=0 \mid X(t_0)=1] + P[K=0 \mid X(t_0)=2]\right\} \\ &= \frac{1}{2}\left\{P[K=0 \mid X(t_0)=1] + 1\right\} \end{aligned}$$

De plus, on peut écrire que

$$P[K=0 \mid X(t_0)=1] = P[N(t_0+S) - N(t_0) = 0] = P[N(S)=0]$$

où $N(t)$ est le nombre d'arrivées dans l'intervalle $[0,t]$ et S est le temps de service d'un client quelconque. En conditionnant sur les valeurs possibles de S, on obtient:

$$P[N(S)=0] = \int_0^\infty P[N(S)=0 \mid S=s] f_S(s) ds$$

Puisque les arrivées des clients et les temps de service sont, par hypothèse, des variables aléatoires indépendantes, on a:

$$\begin{aligned} P[N(S)=0] &= \int_0^\infty P[N(s)=0] \mu e^{-\mu s} ds = \int_0^\infty e^{-\lambda s} \mu e^{-\mu s} ds \\ &= \frac{\mu}{\mu+\lambda} \overset{\lambda=\mu}{=} \frac{1}{2} \end{aligned}$$

Il s'ensuit que

$$P[K=0] = \frac{1}{2}\left(\frac{1}{2}+1\right) = \frac{3}{4}$$

ce qui implique que $P[K=1] = 1/4$ et

$$E[K] = 0 + 1 \times \frac{1}{4} = \frac{1}{4}$$

$\diamondsuit$

Exemple 9.2.4. Écrire les équations d'équilibre pour le système de file d'attente $M/M/1/3$, si l'on suppose que lorsque le serveur finit de servir un client et qu'il y a deux clients en attente, alors il sert ces deux clients en même temps, au taux μ.

Solution. Ici, l'état $X(t)$ du processus ne peut pas être simplement le nombre de clients dans le système à l'instant t. En effet, supposons qu'il y a trois clients dans le système. Le prochain état visité ne sera pas le même si deux clients sont en train d'être servis simultanément ou si deux clients sont en attente. Dans le premier cas, le système fera une transition de l'état 3 à l'état 1, tandis qu'il ira de l'état 3 à l'état 2 dans le deuxième cas. Par conséquent, il faut être plus précis. Soit (m,n) l'état du système s'il y a m client(s) en train d'être servi(s) et n client(s) en attente. Les états possibles sont alors: $(m,0)$, pour $m = 0,1,2$, et $(1,1)$, $(1,2)$ et $(2,1)$. Les équations d'équilibre du système sont les suivantes (voir la figure 9.3):

$$\underline{\text{état } (m,n)}\ \underline{\text{taux de départ de } (m,n)} = \underline{\text{taux d'arrivée à } (m,n)}$$

$$\begin{array}{ll}
(0,0) & \lambda\pi_{(0,0)} = \mu(\pi_{(1,0)} + \pi_{(2,0)}) \\
(1,0) & (\lambda+\mu)\pi_{(1,0)} = \lambda\pi_{(0,0)} + \mu(\pi_{(1,1)} + \pi_{(2,1)}) \\
(1,1) & (\lambda+\mu)\pi_{(1,1)} = \lambda\pi_{(1,0)} \\
(1,2) & \mu\pi_{(1,2)} = \lambda\pi_{(1,1)} \\
(2,0) & (\lambda+\mu)\pi_{(2,0)} = \mu\pi_{(1,2)} \\
(2,1) & \mu\pi_{(2,1)} = \lambda\pi_{(2,0)}
\end{array}$$

Pour obtenir les probabilités limites, on peut résoudre le système d'équations linéaires précédent, sous la condition $\sum_{(m,n)} \pi_{(m,n)} = 1$. On exprime les $\pi_{(m,n)}$ en fonction de $\pi_{(2,1)}$. Pour simplifier, supposons que $\lambda = \mu$. Alors la dernière équation ci-dessus implique que $\pi_{(2,0)} = \pi_{(2,1)}$. Ensuite, on déduit de l'équation de l'état $(2,0)$ que $\pi_{(1,2)} = 2\pi_{(2,1)}$. Il s'ensuit, d'après l'équation de l'état $(1,2)$,

que l'on peut écrire que $\pi_{(1,1)} = 2\pi_{(2,1)}$ également. L'équation de l'état $(1,1)$ permet d'écrire que $\pi_{(1,0)} = 4\pi_{(2,1)}$. Finalement, la première équation donne $\pi_{(0,0)} = 5\pi_{(2,1)}$. Donc, on a:

$$(5+4+2+2+1+1)\pi_{(2,1)} = 1 \quad \Longrightarrow \quad \pi_{(2,1)} = \frac{1}{15}$$

de sorte que

$$\pi_{(0,0)} = \frac{1}{3}, \quad \pi_{(1,0)} = \frac{4}{15}, \quad \pi_{(1,1)} = \pi_{(1,2)} = \frac{2}{15} \quad \text{et} \quad \pi_{(2,0)} = \frac{1}{15}$$

Notons que cette solution satisfait aussi à l'équation de l'état $(1,0)$, dont on ne s'est pas servi pour trouver les probabilités limites.

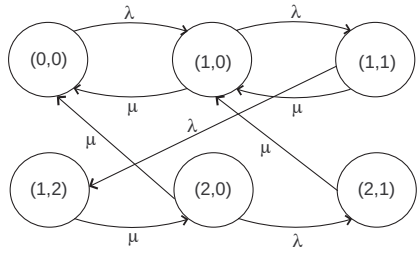

Fig. 9.3. Diagramme de transitions pour le modèle de file d'attente de l'exemple 9.2.4

Le nombre moyen de clients dans le système en régime stationnaire est donné par

$$\sum_{(m,n)} (m+n)\pi_{(m,n)} = 1 \times \pi_{(1,0)} + 2 \times (\pi_{(1,1)} + \pi_{(2,0)}) + 3 \times (\pi_{(1,2)} + \pi_{(2,1)})$$
$$= \frac{4 + 2 \times 3 + 3 \times 3}{15} = \frac{19}{15}$$

et le taux moyen d'entrée des clients dans le système est

$$\lambda_e = \lambda(1 - \pi_{(1,2)} - \pi_{(2,1)}) = \lambda\left(1 - \frac{3}{15}\right) = \frac{4\lambda}{5}$$

$\diamondsuit$

9.3 Systèmes de files d'attente avec deux ou plusieurs serveurs

9.3.1 Modèle $M/M/s$

Supposons que toutes les hypothèses que nous avons faites dans la formulation du modèle d'attente $M/M/1$ sont valables, mais qu'il y a en fait s serveurs dans le système, où $s \in \{2, 3, \ldots\}$. On suppose que les temps de service des s serveurs sont des variables aléatoires $\text{Exp}(\mu)$ indépendantes. Ce modèle est noté $M/M/s$. Le cas particulier où le nombre de serveurs tend vers l'infini sera considéré. De plus, si la capacité du système est finie, on obtient le modèle $M/M/s/c$, lequel est traité dans la prochaine sous-section.

Comme c'est généralement le cas en pratique, on suppose que les clients attendent en une seule file qu'un serveur soit libre (ou ils *prennent un billet* lorsqu'ils entrent dans le système et attendent jusqu'à ce que leur numéro apparaisse sur un écran lumineux). Cela signifie que s'il y a au plus s clients dans le système, alors ils sont tous en train d'être servis, ce qui n'est pas nécessairement le cas si on suppose qu'une file d'attente se forme devant chacun des serveurs. De plus, comme il était implicite dans la section précédente, la *politique de service* est celle de *premier arrivé, premier servi* (notée $FIFO$, pour *First In, First Out*, en anglais).

Remarque. Dans les exemples et les exercices, nous modifions souvent le modèle de base $M/M/s$. Par exemple, nous pouvons supposer que les serveurs ne servent pas nécessairement tous au même taux μ, ou que la politique de service est différente de celle par défaut (c'est-à-dire $FIFO$), etc.

Soit $X(t)$ le nombre de clients dans le système à l'instant $t \geq 0$. Le processus stochastique $\{X(t), t \geq 0\}$ est une chaîne de Markov à temps continu. Le processus d'arrivée est un processus de Poisson de taux $\lambda > 0$. De plus, même s'il y a au moins deux serveurs, puisque les clients sont servis un à la fois et que les temps de service sont des variables aléatoires exponentielles (donc, *continues*), deux clients (ou plus) ne peuvent pas quitter le système exactement au même instant. Il s'ensuit que $\{X(t), t \geq 0\}$ est un processus de naissance et de mort. Les taux d'arrivée λ_k sont tous égaux à λ, et les taux de départ μ_k sont donnés par

$$\mu_k = \begin{cases} k\mu \text{ si } k = 1, \ldots, s-1 \\ s\mu \text{ si } k = s, s+1, \ldots \end{cases}$$

En effet, lorsqu'il y a k clients en train d'être servis simultanément, le temps requis pour qu'un départ ait lieu est le minimum de k variables aléatoires $\text{Exp}(\mu)$

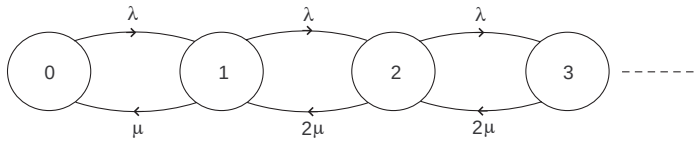

Fig. 9.4. Diagramme de transitions pour le modèle $M/M/2$

indépendantes. On sait que ce minimum présente une distribution exponentielle de paramètre $\mu + \cdots + \mu = k\mu$ (voir la remarque qui suit la proposition 8.2.1).

On déduit de ce qui précède que les équations d'équilibre du système de file d'attente $M/M/s$ sont (voir la figure 9.4 pour le cas où $s = 2$):

$$\underline{\text{état } j} \qquad \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc} 0 & \lambda\pi_0 = \mu\pi_1 \\ k \in \{1, \ldots, s-1\} & (\lambda + k\mu)\pi_k = (k+1)\mu\pi_{k+1} + \lambda\pi_{k-1} \\ k \in \{s, s+1, \ldots\} & (\lambda + s\mu)\pi_k = s\mu\pi_{k+1} + \lambda\pi_{k-1} \end{array}$$

Pour résoudre ce système d'équations linéaires, sous la condition $\sum_{k=0}^{\infty} \pi_k = 1$, on fait appel au théorème 9.1.1. D'abord, on calcule

$$\Pi_k = \frac{\lambda \times \lambda \times \cdots \times \lambda}{\mu \times 2\mu \times \cdots \times k\mu} = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k = \frac{1}{k!}\rho^k \quad \text{pour } k = 1, 2, \ldots, s$$

Dans le cas où $k = s+1, s+2, \ldots$, on trouve que

$$\Pi_k = \frac{\rho^k}{s!s^{k-s}}$$

Ensuite, la somme $\sum_{k=1}^{\infty} \Pi_k$ converge si et seulement si

$$\sum_{k=s+1}^{\infty} \Pi_k < \infty \iff \sum_{k=s+1}^{\infty} \frac{\rho^k}{s!s^{k-s}} < \infty$$

$$\iff \frac{s^s}{s!} \sum_{k=s+1}^{\infty} \left(\frac{\rho}{s}\right)^k < \infty \iff \rho < s$$

ce qui, encore une fois, équivaut à dire que le taux d'arrivée des clients doit être inférieur à leur taux de service (maximal).

Maintenant, puisque les taux d'arrivée et de départ sont strictement positifs, le processus de naissance et de mort $\{X(t),\ t \geq 0\}$ est irréductible et les probabilités limites peuvent effectivement être obtenues en faisant appel au théorème 9.1.1. On trouve (après un peu de travail) que

$$\pi_0 = \left[\sum_{k=0}^{s-1} \frac{\rho^k}{k!} + \frac{\rho^s}{s!}\frac{s}{(s-\rho)}\right]^{-1} \quad \text{si } \rho < s \tag{9.18}$$

On peut alors écrire que

$$\pi_k = \frac{\rho^k}{k!}\pi_0 \quad \text{si } k = 1, \ldots, s \tag{9.19}$$

et

$$\pi_k = \frac{\rho^k}{s!s^{k-s}}\pi_0 \quad \text{si } k = s+1, s+2, \ldots \tag{9.20}$$

Obtenir les quantités $\bar{N}$ et $\bar{T}$ requiert un peu d'effort. D'abord, parce que les taux de service sont, par hypothèse, tous égaux à μ, on peut écrire que $\bar{S} = 1/\mu$. Il s'ensuit, à partir de la formule de Little (avec $\lambda_e = \lambda$) que

$$\bar{N}_S = \lambda\bar{S} = \rho$$

Ensuite, on peut montrer que

$$\bar{N}_Q = \frac{\rho^{s+1}}{s!}\frac{s}{(s-\rho)^2}\pi_0$$

d'où l'on déduit que

$$\bar{N} = \frac{\rho^{s+1}}{s!}\frac{s}{(s-\rho)^2}\pi_0 + \rho$$

Finalement, on a:

$$\bar{Q} = \frac{\bar{N}_Q}{\lambda} = \frac{\rho^{s+1}}{\lambda s!}\frac{s}{(s-\rho)^2}\pi_0 \quad \text{et} \quad \bar{T} = \bar{Q} + \frac{1}{\mu}$$

Remarque. Dans certaines applications, le nombre de *serveurs* est infini. Par exemple, supposons que les *clients* sont des personnes qui se promènent dans un parc public et que leur *temps de service* est la quantité aléatoire de temps qu'elles passent dans le parc. Parce que les personnes n'ont pas à attendre d'être

servies, cette situation correspond au cas où s est l'infini. On peut écrire (voir l'équation (9.18)) que

$$\lim_{s\to\infty} \pi_0 = \left(\sum_{k=0}^{\infty} \frac{\rho^k}{k!} \right)^{-1} = e^{-\rho}$$

de sorte que

$$\lim_{s\to\infty} \pi_k = \frac{\rho^k}{k!} e^{-\rho} \quad \text{pour } k = 1, 2, \ldots \tag{9.21}$$

De là, si $K \sim \text{Poi}(\rho)$, on peut affirmer que les probabilités limites pour le modèle $M/M/\infty$ sont données par

$$\pi_k = P[K = k] \quad \text{pour } k = 0, 1, \ldots \tag{9.22}$$

Il s'ensuit que $\bar{N} = \rho$. Finalement, parce que les clients n'attendent jamais d'être servis, on a que $\bar{N}_Q = \bar{Q} = 0$, de sorte que $\bar{N}_S = \bar{N} = \rho$ et $\bar{T} = \bar{S} = 1/\mu$.

Maintenant, une autre quantité d'intérêt est la probabilité que tous les serveurs soient occupés lorsque le système est en régime stationnaire. On symbolise cette probabilité par π_b. On peut montrer que

$$\pi_b \equiv \sum_{k\geq s} \pi_k = \frac{s\rho^s}{s!(s-\rho)} \pi_0 \quad \text{si } \rho < s$$

Il est possible d'exprimer en fonction de π_b la fonction de répartition du temps Q qu'un client quelconque passe à faire la queue. La variable aléatoire Q est de type mixte. On a que $P[Q = 0] = 1 - \pi_b$, car $1 - \pi_b$ est la probabilité que le client en question arrive alors qu'il y a moins de s clients déjà présents dans le système. En général, on peut montrer que, si $\rho < s$,

$$P[Q \leq t] = 1 - \pi_b e^{(\rho - s)t} \quad \text{pour } t \geq 0$$

Exemple 9.3.1. Considérons le modèle d'attente $M/M/2$. Supposons que les deux serveurs sont occupés. Calculer la probabilité que les temps de service des deux clients en train d'être servis diffèrent d'au plus une unité de temps.

Solution. Soit S_i le temps de service du client en train d'être servi par le serveur n$^{\text{o}}$ i, pour $i = 1, 2$. Les variables S_i sont des variables aléatoires $\text{Exp}(\mu)$ indépendantes. Par symétrie, on peut écrire que

$$
\begin{aligned}
P[|S_1 - S_2| \le 1] &= 2P[S_1 \le S_2 \le S_1 + 1] \\
&= 2\int_0^\infty \int_{s_1}^{s_1+1} \mu e^{-\mu s_1} \mu e^{-\mu s_2} ds_2 ds_1 \\
&= 2\int_0^\infty \mu e^{-\mu s_1} \left\{ \int_{s_1}^{s_1+1} \mu e^{-\mu s_2} ds_2 \right\} ds_1 \\
&= 2\int_0^\infty \mu e^{-\mu s_1} \left\{ -e^{-\mu s_2} \Big|_{s_1}^{s_1+1} \right\} ds_1 \\
&= 2\left(1 - e^{-\mu}\right) \int_0^\infty \mu e^{-2\mu s_1} ds_1 = 2\left(1 - e^{-\mu}\right) \frac{1}{2}
\end{aligned}
$$

Donc, la probabilité requise est égale à $1 - e^{-\mu}$.

Remarque. Si on n'utilise pas la symétrie, on doit calculer deux intégrales doubles (voir la figure 9.5):

$$
\begin{aligned}
P[|S_1 - S_2| \le 1] = &\int_0^1 \int_0^{s_1+1} \mu e^{-\mu s_1} \mu e^{-\mu s_2} ds_2 ds_1 \\
&+ \int_1^\infty \int_{s_1-1}^{s_1+1} \mu e^{-\mu s_1} \mu e^{-\mu s_2} ds_2 ds_1
\end{aligned}
$$

♢

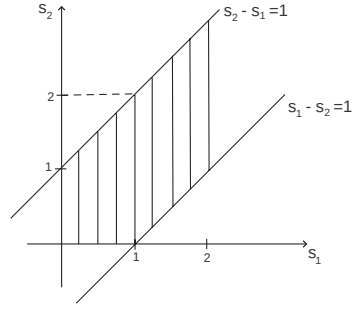

Fig. 9.5. Figure pour l'exemple 9.3.1

Exemple 9.3.2. Supposons que $\lambda = 2\mu$ dans un système d'attente $M/M/3$.

(a) Quelle est la variance du nombre de clients dans le système à un instant t_0 assez grand pour que le système soit en régime stationnaire, étant donné que personne ne fait la queue à l'instant t_0?

(b) Sachant que les trois serveurs sont occupés à l'instant t_0, quelle est la probabilité que personne ne fasse la queue?

Solution. (a) D'abord, t_0 étant assez grand, on peut écrire (voir l'équation (9.19)) que

$$\pi_k^* := P[X(t_0) = k \mid X(t_0) \leq 3] = \frac{\pi_k}{\pi_0 + \pi_1 + \pi_2 + \pi_3} = \frac{\rho^k/k!}{\sum_{j=0}^{3} \rho^j/j!}$$
$$\overset{\rho=2}{=} \left(\frac{3}{19}\right)\left(\frac{2^k}{k!}\right) \quad \text{pour } k = 0, 1, 2, 3$$

C'est-à-dire que $\pi_0^* = 3/19$, $\pi_1^* = \pi_2^* = 6/19$ et $\pi_3^* = 4/19$.

Remarque. Notons que parce que les probabilités limites π_k sont exprimées en fonction de π_0, on n'a pas eu besoin de calculer la valeur de π_0 pour obtenir π_k. En fait, on a:

$$\pi_0 = \left[\sum_{k=0}^{2} \frac{2^k}{k!} + \frac{2^3}{3!}\frac{3}{(3-2)}\right]^{-1} = \frac{1}{9}$$

Ensuite, on calcule

$$E[X(t_0) \mid X(t_0) \leq 3] = \frac{1}{19}\left(6 + 2 \times 6 + 3 \times 4\right) = \frac{30}{19}$$

et

$$E[X^2(t_0) \mid X(t_0) \leq 3] = \frac{1}{19}\left(6 + 2^2 \times 6 + 3^2 \times 4\right) = \frac{66}{19}$$

de sorte que

$$\mathrm{VAR}[X(t_0) \mid X(t_0) \leq 3] = \frac{66}{19} - \left(\frac{30}{19}\right)^2 = \frac{354}{361}$$

(b) On cherche

$$p := P[X(t_0) = 3 \mid X(t_0) \geq 3] = \frac{P[X(t_0) = 3]}{P[X(t_0) \geq 3]} = \frac{P[X(t_0) = 3]}{1 - P[X(t_0) \leq 2]}$$
$$\overset{(9.19)}{=} \frac{(4/3)\pi_0}{1 - (\pi_0 + 2\pi_0 + 2\pi_0)}$$

Ainsi, on a besoin ici de la valeur explicite de π_0. En utilisant la remarque précédente, on peut écrire que

$$p = \frac{4/3}{\pi_0^{-1} - (1 + 2 + 2)} = \frac{4/3}{9 - (1 + 2 + 2)} = \frac{1}{3}$$

♢

9.3.2 Modèle $M/M/s/c$

Même si la capacité, c, d'un système $M/M/s$ est finie, le processus stochastique $\{X(t), t \geq 0\}$ demeure un processus de naissance et de mort. Par conséquent, on peut faire appel au théorème 9.1.1 pour obtenir les probabilités limites du processus. Le cas pour lequel $c = s$ est particulièrement important. Le modèle $M/M/s/s$ est un cas particulier des systèmes appelés *systèmes avec perte*, parce que lorsque tous les serveurs sont occupés, les clients qui arrivent ne peuvent pas (ou ne veulent pas) entrer dans le système. Ainsi, ils sont *perdus*. Un exemple de système de ce type est un parc de stationnement. Dans ce cas, les places de stationnement libres sont les *serveurs*, et lorsque le parc de stationnement est plein, les conducteurs qui arrivent doivent aller garer leur voiture ailleurs.

Nous allons maintenant calculer les probabilités limites pour le modèle $M/M/s/s$. Les équations d'équilibre du système sont les suivantes:

$$\begin{array}{cc} \underline{\text{état } j} & \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j} \\ 0 & \lambda\pi_0 = \mu\pi_1 \\ k \in \{1, \ldots, s-1\} & (\lambda + k\mu)\pi_k = (k+1)\mu\pi_{k+1} + \lambda\pi_{k-1} \\ s & s\mu\pi_s = \lambda\pi_{s-1} \end{array}$$

Le processus de naissance et de mort $\{X(t), t \geq 0\}$, où $X(t)$ représente le nombre de clients dans le système à l'instant t, est irréductible. De plus, la capacité du système étant finie, les probabilités limites existent pour toutes les valeurs permises des paramètres λ et μ. On calcule

$$\Pi_k = \frac{\lambda \times \lambda \times \cdots \times \lambda}{\mu \times 2\mu \times \cdots \times k\mu} = \frac{\rho^k}{k!} \quad \text{pour } k = 1, 2, \ldots, s$$

Il s'ensuit que

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{s} \rho^k/k!} = \left(\sum_{k=0}^{s} \frac{\rho^k}{k!} \right)^{-1} \tag{9.23}$$

et

$$\pi_j = \frac{\rho^j}{j!}\pi_0 \quad \text{pour } j = 1, \ldots, s \tag{9.24}$$

Remarques.
(i) Si s tend vers l'infini, on devrait retrouver les résultats obtenus pour la file d'attente $M/M/\infty$ dans la sous-section précédente (voir l'équation (9.21)). En effet, on a:

$$\lim_{s\to\infty} \pi_0 = \left(\sum_{k=0}^{\infty} \frac{\rho^k}{k!} \right)^{-1} = (e^{\rho})^{-1} = e^{-\rho}$$

de sorte que

$$\lim_{s\to\infty} \pi_j = \frac{\rho^j}{j!} e^{-\rho} \quad \text{pour } j = 1, 2, \ldots$$

(ii) La probabilité π_b que tous les serveurs soient occupés est donnée par

$$\pi_b = \pi_s = \frac{\rho^s}{s!}\pi_0 = \frac{\rho^s/s!}{\sum_{j=0}^{s} \rho^j/j!} \tag{9.25}$$

Cette formule est appelée *formule d'Erlang*.

(iii) Puisque $\rho := \lambda/\mu = \lambda E[S]$, les formules des probabilités limites peuvent être réécrites comme suit:

$$\pi_0 = \left(\sum_{k=0}^{s} \frac{(\lambda E[S])^k}{k!} \right)^{-1} \quad \text{et} \quad \pi_j = \frac{(\lambda E[S])^j}{j!}\pi_0 \quad \text{pour } j = 1, \ldots, s \tag{9.26}$$

Un résultat très intéressant permet d'affirmer que les formules précédentes sont valables même si la variable aléatoire S ne présente pas une distribution exponentielle, pourvu qu'elle soit non négative. Par exemple, S pourrait présenter une distribution uniforme sur l'intervalle $(0, 1)$, ou une distribution gamma, etc. C'est-à-dire que l'équation (9.26) est valable pour le modèle $M/G/s/s$ (appelé *système avec perte d'Erlang*), où G signifie *général*.

Pour compléter cette sous-section, nous allons calculer les diverses quantités d'intérêt, ce qui s'avère facile dans ce cas, car $Q \equiv 0$. Il s'ensuit immédiatement que $\bar{Q} = \bar{N}_Q = 0$. Puisque $\bar{S} = 1/\mu$, comme auparavant, on peut écrire que

$$\bar{T} = \bar{S} = \frac{1}{\mu}$$

et

$$\bar{N} = \bar{N}_S = \lambda_e \bar{S} = \lambda(1 - \pi_s)\frac{1}{\mu} = (1 - \pi_s)\rho$$

Exemple 9.3.3. Les équations d'équilibre du système de file d'attente $M/M/2/3$ avec $\lambda = 2\mu$ sont:

$$\underline{\text{état } j}\ \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc} 0 & 2\mu\pi_0 = \mu\pi_1 \\ 1 & (2\mu+\mu)\pi_1 = 2\mu\pi_0 + (2\times\mu)\pi_2 \\ 2 & (2\mu+2\times\mu)\pi_2 = 2\mu\pi_1 + (2\times\mu)\pi_3 \\ 3 & (2\times\mu)\pi_3 = 2\mu\pi_2 \end{array}$$

C'est-à-dire que

$$\begin{aligned} 2\pi_0 &\overset{(0)}{=} \pi_1 \\ 3\pi_1 &\overset{(1)}{=} 2\pi_0 + 2\pi_2 \\ 2\pi_2 &\overset{(2)}{=} \pi_1 + \pi_3 \\ \pi_3 &\overset{(3)}{=} \pi_2 \end{aligned}$$

Il est facile de résoudre ce système d'équations linéaires. Les équations des états 2 et 3 impliquent que $\pi_1 = \pi_2 = \pi_3$. Alors, en utilisant l'équation de l'état 0, on peut écrire que

$$\pi_0 + 2\pi_0 + 2\pi_0 + 2\pi_0 = 1 \quad \Longrightarrow \quad \pi_0 = \frac{1}{7} \quad \text{et} \quad \pi_1 = \pi_2 = \pi_3 = \frac{2}{7}$$

Finalement, on vérifie immédiatement que cette solution satisfait également à l'équation de l'état 1.

Dans le cas du système de file d'attente $M/M/2/2$ (avec $\lambda = 2\mu$), on déduit des équations (9.23) et (9.24) que

$$\pi_0 = \left(\sum_{k=0}^{2} \frac{2^k}{k!}\right)^{-1} = (1+2+2)^{-1} = \frac{1}{5} \quad \text{et} \quad \pi_1 = \pi_2 = \frac{2}{5}$$

9.4 Exercices du chapitre 9

Exercices résolus

Question n^{o} 1

Un système est constitué de n composants fonctionnant indépendamment les uns des autres et possédant une durée de vie qui présente une distribution exponentielle de paramètre μ_k, pour $k = 1, \ldots, n$. Lorsque le système cesse de fonctionner, les composants en panne sont remplacés par des composants neufs. Soit $N(t)$ le nombre de pannes du système dans l'intervalle $[0, t]$. Le processus stochastique $\{N(t), t \geq 0\}$ est-il une chaîne de Markov à temps continu si les composants sont placés (a) en série? (b) en parallèle? (c) en redondance passive?

Solution. (a) Soit T_k, pour $k = 1, \ldots, n$, la durée de vie du composant n^{o} k. Si les composants sont placés en série, alors, par la propriété de non-vieillissement de la distribution exponentielle, on peut exprimer le temps T entre deux pannes consécutives du système comme suit: $T = \min\{T_1, \ldots, T_n\}$. À partir de la remarque qui suit la proposition 8.2.1, on déduit que $T \sim \text{Exp}(\mu_1 + \cdots + \mu_n)$. De là, on peut affirmer que $\{N(t), t \geq 0\}$ *est* une chaîne de Markov à temps continu. En fait, c'est un processus de Poisson de taux $\lambda = \mu_1 + \cdots + \mu_n$.

(b) Lorsque les composants sont placés en parallèle, on a: $T = \max\{T_1, \ldots, T_n\}$. Or, le maximum $T_{1,2}$ de deux variables aléatoires exponentielles indépendantes ne présente pas une distribution exponentielle. En effet, on peut écrire (voir l'exemple 8.2.2) que

$$P[T_{1,2} \leq t] = \left(1 - e^{-\mu_1 t}\right)\left(1 - e^{-\mu_2 t}\right) \quad \text{pour } t \geq 0$$

de sorte que

$$f_{T_{1,2}}(t) = \frac{d}{dt} P[T_{1,2} \leq t] = \mu_1 e^{-\mu_1 t} + \mu_2 e^{-\mu_2 t} - (\mu_1 + \mu_2) e^{-(\mu_1 + \mu_2)t} \quad \text{pour } t \geq 0$$

Parce qu'on ne peut pas écrire $f_{T_{1,2}}(t)$ sous la forme

$$f_{T_{1,2}}(t) = \lambda e^{-\lambda t} \quad \text{pour } t \geq 0$$

pour un certain $\lambda > 0$, on doit conclure que $T_{1,2}$ ne présente pas une distribution exponentielle. Par extension, T n'est pas une variable aléatoire exponentielle non plus, de sorte que $\{N(t),\, t \geq 0\}$ *n'est pas* une chaîne de Markov à temps continu.

(c) Dans le cas où les composants sont placés en redondance passive, la variable aléatoire T ne présente pas une distribution exponentielle non plus (pour $n \geq 2$). En effet, si $\mu_k \equiv \mu$, alors (voir l'équation (4.27)) $T \sim \mathrm{G}(n, \mu)$. Parce que T ne présente pas une distribution exponentielle dans ce cas particulier, elle ne peut pas présenter une distribution exponentielle pour des μ_k quelconques. Donc, $\{N(t), t \geq 0\}$ *n'est pas* une chaîne de Markov à temps continu.

Question n° 2

Le processus de naissance pur particulier connu sous le nom de *processus de Yule* est tel que $\lambda_n = n\lambda$, pour $n = 0, 1, \ldots$ On peut montrer que

$$p_{i,j}(t) = \binom{j-1}{i-1} e^{-i\lambda t}(1 - e^{-\lambda t})^{j-i} \quad \text{pour } j \geq i \geq 1$$

Quelle est l'espérance mathématique de $X(t)$, étant donné que $X(0) = i > 0$?

Solution. On peut écrire que $X(t)$, étant donné que $X(0) = i$, présente une distribution binomiale négative de paramètres $r = i$ et $p = e^{-\lambda t}$ (voir l'équation (3.9)). À partir du tableau 3.1, page 107, on déduit que l'espérance mathématique de $X(t)$ est donnée par

$$E[X(t) \mid X(0) = i] = \frac{r}{p} = ie^{\lambda t}$$

Pour justifier ce résultat, notons que

$$p_{1,j}(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{j-1} \quad \text{pour } j \geq 1$$

C'est-à-dire que

$$P[X(t) = j \mid X(0) = 1] = P[\text{Géom}(p := e^{-\lambda t}) = j] \quad \text{pour } j = 1, 2, \ldots$$

Maintenant, lorsque $X(0) = i \geq 1$, on peut représenter $X(t)$ comme la somme de i variables aléatoires géométriques (indépendantes) de paramètre commun p. Alors, par linéarité de l'espérance mathématique, on déduit que

$$E[X(t) \mid X(0) = i] = i \times E[\text{Géom}(p := e^{-\lambda t})] = i\left(\frac{1}{e^{-\lambda t}}\right) = ie^{\lambda t}$$

Question n° 3

Soit $\{X(t), t \geq 0\}$ un processus de naissance et de mort dont l'espace des états est $\{0, 1, 2\}$ et dont les taux de naissance et de mort sont donnés par

$$\lambda_0 = \lambda, \quad \lambda_1 = 2\lambda \quad \text{et} \quad \mu_1 = \mu, \quad \mu_2 = 2\mu$$

Trouver les probabilités limites du processus à partir de ses équations d'équilibre.

Solution. Les équations d'équilibre du système sont les suivantes:

$$\underline{\text{état } j} \; \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc} 0 & \lambda\pi_0 = \mu\pi_1 \\ 1 & (2\lambda+\mu)\pi_1 = \lambda\pi_0 + 2\mu\pi_2 \\ 2 & 2\mu\pi_2 = 2\lambda\pi_1 \end{array}$$

On déduit de la première et de la troisième équation que

$$\pi_1 = \frac{\lambda}{\mu}\pi_0 \quad \text{et} \quad \pi_2 = \frac{\lambda}{\mu}\pi_1 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$$

De là, on peut écrire que

$$\pi_0 + \frac{\lambda}{\mu}\pi_0 + \left(\frac{\lambda}{\mu}\right)^2 \pi_0 = 1 \quad \Longrightarrow \quad \pi_0 = \left[1 + \frac{\lambda}{\mu} + \left(\frac{\lambda}{\mu}\right)^2\right]^{-1}$$

d'où l'on obtient les valeurs de π_1 et π_2.

Question n$^{\text{o}}$ 4

Trouver les probabilités limites d'un modèle d'attente $M/M/1$ à un instant t_0 (assez grand), étant donné qu'il y a soit deux, trois ou quatre clients dans le système à cet instant. Sous la même condition, quelle est l'espérance mathématique du temps que le premier client qui entre dans le système après t_0 passera à faire la queue, si on suppose que le client qui était en train d'être servi à l'instant t_0 est encore présent lorsque le nouveau client arrive?

Solution. Soit $\pi_k^* = P[X(t_0) = k \mid X(t_0) \in \{2,3,4\}]$. On peut écrire que

$$\pi_k^* = \frac{\pi_k}{\pi_2 + \pi_3 + \pi_4} = \frac{\rho^k}{\rho^2 + \rho^3 + \rho^4} = \frac{\rho^{k-2}}{1 + \rho + \rho^2} \quad \text{pour } k = 2,3,4$$

Ensuite, soit Q^* le temps d'attente du nouveau client. Par la propriété de non-vieillissement de la distribution exponentielle, on peut écrire que

$$E[Q^* \mid X(t_0) \in \{2,3,4\}] = \sum_{k=2}^{4} E[Q^* \mid \{X(t_0) = k\} \cap \{X(t_0) \in \{2,3,4\}\}] \times P[X(t_0) = k \mid X(t_0) \in \{2,3,4\}] = \sum_{k=2}^{4} \left(\frac{k}{\mu}\right) \pi_k^* = \frac{1}{\mu}\left(\frac{2+3\rho+4\rho^2}{1+\rho+\rho^2}\right)$$

Question n° 5

On suppose que le serveur dans un système de file d'attente $M/M/1$ travaille deux fois plus rapidement lorsqu'il y a au moins trois clients dans le système, de sorte que $\mu[X(t)] = \mu$ si $X(t) = 1$ ou 2, et $\mu[X(t)] = 2\mu$ si $X(t) \geq 3$. Écrire les équations d'équilibre de ce système. Quelle est la condition de l'existence des probabilités limites?

Solution. Les équations d'équilibre sont:

état j	taux de départ de j = taux d'arrivée à j
0	$\lambda\pi_0 = \mu\pi_1$
1	$(\lambda+\mu)\pi_1 = \lambda\pi_0 + \mu\pi_2$
2	$(\lambda+\mu)\pi_2 = \lambda\pi_1 + 2\mu\pi_3$
$n \in \{3,4,\ldots\}$	$(\lambda+2\mu)\pi_n = \lambda\pi_{n-1} + 2\mu\pi_{n+1}$

Ensuite, on calcule

$$\Pi_k = \frac{\lambda\lambda\lambda\cdots\lambda}{\mu\mu(2\mu)\cdots(2\mu)} = \left(\frac{\lambda}{\mu}\right)^2 \left(\frac{\lambda}{2\mu}\right)^{k-2} \quad \text{pour } k = 2,3,\ldots$$

De là, on déduit que la somme $\sum_{k=1}^{\infty} \Pi_k$ converge si et seulement si

$$\sum_{k=2}^{\infty} \Pi_k < \infty \iff \sum_{k=2}^{\infty} \left(\frac{\lambda}{\mu}\right)^2 \left(\frac{\lambda}{2\mu}\right)^{k-2} < \infty$$
$$\iff \left(\frac{\lambda}{\mu}\right)^2 \sum_{k=2}^{\infty} \left(\frac{\lambda}{2\mu}\right)^{k-2} < \infty \iff \frac{\lambda}{2\mu} < 1$$

comme on aurait pu le deviner.

Question n° 6

On considère le système de file d'attente $M/M/1/2$ en régime stationnaire. On suppose qu'un départ a eu lieu à l'instant t_0 et que les deux arrivées suivantes,

à partir de t_0, se sont produites à $t = t_0+1$ et $t = t_0+2$. Quelle est la probabilité que le client qui est arrivé à l'instant $t_0 + 2$ ait pu entrer dans le système?

Solution. Les probabilités limites du système sont données par (voir l'équation (9.17))

$$\pi_j = \frac{\rho^j(1-\rho)}{1-\rho^3} \quad \text{pour } j = 0, 1, 2$$

Soit F l'événement suivant: *le client qui est arrivé à l'instant* $t_0 + 2$ *a pu entrer dans le système*. On peut écrire que

$$\begin{aligned} P[F] &= P[F \mid X(t_0^-) = 1]P[X(t_0^-) = 1 \mid X(t_0^-) \in \{1, 2\}] \\ &\quad + P[F \mid X(t_0^-) = 2]P[X(t_0^-) = 2 \mid X(t_0^-) \in \{1, 2\}] \\ &= 1 \times \frac{\pi_1}{\pi_1 + \pi_2} + P[F \mid X(t_0^-) = 2]\frac{\pi_2}{\pi_1 + \pi_2} \end{aligned}$$

Maintenant, étant donné que le système était plein lorsqu'un départ a eu lieu à l'instant t_0, le client qui est arrivé à $t_0 + 2$ a pu entrer dans le système si et seulement si celui qui a commencé à être servi à t_0 a quitté le système avant l'instant $t_0 + 2$. C'est-à-dire que si $S \sim \text{Exp}(\mu)$,

$$P[F \mid X(t_0^-) = 2] = P[S < 2] = 1 - e^{-2\mu}$$

De là, la probabilité requise est

$$P[F] = \frac{\rho}{\rho + \rho^2} + \left(1 - e^{-2\mu}\right)\frac{\rho^2}{\rho + \rho^2} = \frac{1}{1+\rho}\left[1 + \rho\left(1 - e^{-2\mu}\right)\right]$$

Question n$^{\text{o}}$ 7

On suppose que le serveur dans un système de file d'attente $M/M/1/3$ décide de travailler deux fois plus rapidement, de façon à augmenter les profits du système. Cependant, après un certain temps, le taux d'arrivée des clients diminue de λ à $\lambda/2$ à cause du piètre service. Soit $\lambda = \mu$. Si chaque client qui entre vraiment dans le système paie x \$, quel montant d'argent moyen le système gagne-t-il par unité de temps lorsque le taux de service est μ? Est-ce préférable pour le serveur de servir au taux μ ou au taux 2μ?

Solution. Lorsque $\lambda_a = \lambda$ et que le taux de service est $\mu = \lambda$, les probabilités limites du système sont (voir l'équation (9.16)):

$$\pi_j = \frac{1}{4} \quad \text{pour } j = 0, 1, 2, 3$$

Le taux moyen d'entrée des clients dans le système est $\lambda_e = \lambda(1 - \pi_3) = 3\lambda/4$, de sorte que le montant d'argent moyen que le système gagne par unité de temps est égal à $x\,\$ \times 3\lambda/4$.

Ensuite, si $\lambda^* = \lambda/2$ et $\mu^* = 2\mu$, alors $\rho^* = \rho/4 = 1/4$ et (voir l'équation (9.17))

$$\pi_j^* = \frac{(1/4)^j[1 - (1/4)]}{1 - (1/4)^4} \quad \text{pour } j = 0, 1, 2, 3$$

Il s'ensuit que

$$\lambda_e^* = \lambda^*(1 - \pi_3^*) = \frac{\lambda}{2}\left(1 - \frac{3}{255}\right) = \frac{126\lambda}{255}$$

et, parce que les clients qui entrent dans le système paient le même montant qu'auparavant, on conclut qu'il est plus avantageux pour le serveur de servir au taux μ. En effet, on a:

$$x\,\$ \times \frac{3\lambda}{4} > x\,\$ \times \frac{126\lambda}{255}$$

Question n$^{\text{o}}$ 8

Soit $X(t)$ le nombre de clients à l'instant t dans un modèle d'attente $M/M/2$. On suppose que $X(t_0) \geq 2$, et soit τ_i l'instant où le client en train d'être servi par le serveur n$^{\text{o}}$ i, pour $i = 1, 2$, a quitté le système. Calculer la probabilité $P[\tau_2 \leq \tau_1 + 1]$.

Solution. Par la propriété de non-vieillissement de la distribution exponentielle, on peut écrire que $W_i := \tau_i - t_0$ est une variable aléatoire qui présente une distribution exponentielle de paramètre μ. En effet, on a:

$$P[W_i > t] = P[S_i - (t_0 - t_i) > t \mid S_i > t_0 - t_i] \quad \text{pour } i = 1, 2$$

où S_i est le temps de service du client en train d'être servi par le serveur n$^{\text{o}}$ i, et $t_i > 0$ est son instant (connu) d'arrivée. Puisque S_i est une variable aléatoire $\text{Exp}(\mu)$, on peut écrire que

$$\begin{aligned} P[S_i - (t_0 - t_i) > t \mid S_i > t_0 - t_i] &= P[S_i > t + (t_0 - t_i) \mid S_i > t_0 - t_i] \\ &= P[S_i > t] = e^{-\mu t} \end{aligned}$$

De là, par indépendance, la probabilité requise est donnée par

$$\begin{aligned}
&P[\tau_2 - t_0 \leq \tau_1 - t_0 + 1] = P[W_2 \leq W_1 + 1] \\
&= P[S_2 \leq S_1 + 1] = \int_0^\infty \int_0^{s_1+1} \mu e^{-\mu s_1} \mu e^{-\mu s_2} ds_2 ds_1 \\
&= \int_0^\infty (-1)\mu e^{-\mu(s_1+s_2)} \Big|_{s_2=0}^{s_2=s_1+1} ds_1 = \int_0^\infty \mu \left[e^{-\mu s_1} - e^{-\mu(2s_1+1)} \right] ds_1 \\
&= 1 - \frac{1}{2} e^{-\mu}
\end{aligned}$$

Remarque. Puisque, par symétrie et continuité, $P[S_1 < S_2] = 1/2$, on déduit que

$$P[S_1 \leq S_2 \leq S_1 + 1] = 1 - \frac{1}{2} e^{-\mu} - \frac{1}{2} = \frac{1}{2} \left(1 - e^{-\mu}\right)$$

d'où l'on retrouve le résultat de l'exemple 9.3.1.

Question n° 9

Écrire les équations d'équilibre pour le système de file d'attente $M/M/2/3$, si l'on suppose que le temps de service du serveur n° i présente une distribution exponentielle de paramètre μ_i, pour $i = 1, 2$. C'est-à-dire que les deux serveurs ne travaillent pas nécessairement à la même vitesse. On suppose que, lorsque le système est vide, un client qui arrive se dirige vers le serveur n° 1 avec une probabilité de 1. En fonction des probabilités limites du processus, quel est le temps moyen qu'un client qui entre dans le système (en régime stationnaire) passe dans ce système?

Solution. Parce que les taux de service ne sont pas nécessairement égaux, $X(t)$ ne peut pas être simplement le nombre de clients dans le système à l'instant t. On définit les états:

0 : le système est vide
1_1 : seul le serveur n° 1 est occupé
1_2 : seul le serveur n° 2 est occupé
2 : les deux serveurs sont occupés et personne n'attend
3 : les deux serveurs sont occupés et quelqu'un attend

On a:

$$\underline{\text{état } j}\ \underline{\text{taux de départ de } j} = \underline{\text{taux d'arrivée à } j}$$

$$\begin{array}{cc}
0 & \lambda\pi_0 = \mu_1\pi_{1_1} + \mu_2\pi_{1_2} \\
1_1 & (\lambda+\mu_1)\pi_{1_1} = \lambda\pi_0 + \mu_2\pi_2 \\
1_2 & (\lambda+\mu_2)\pi_{1_2} = \mu_1\pi_2 \\
2 & (\lambda+\mu_1+\mu_2)\pi_2 = \lambda(\pi_{1_1}+\pi_{1_2}) + (\mu_1+\mu_2)\pi_3 \\
3 & (\mu_1+\mu_2)\pi_3 = \lambda\pi_2
\end{array}$$

La valeur de $\bar{N}$ est donnée par

$$\bar{N} = \pi_{1_1} + \pi_{1_2} + 2\pi_2 + 3\pi_3$$

De plus, le taux moyen d'entrée des clients est $\lambda_e = \lambda(1-\pi_3)$. Il s'ensuit, à partir de la formule de Little, que

$$\bar{T} = \frac{\bar{N}}{\lambda_e} = \frac{\pi_{1_1} + \pi_{1_2} + 2\pi_2 + 3\pi_3}{\lambda(1-\pi_3)}$$

Question n$^{\text{o}}$ 10

Des automobilistes arrivent selon un processus de Poisson de taux λ à une station-service ayant deux distributeurs d'essence, mais aucun espace d'attente pour les voitures. On suppose que le temps de service est une variable aléatoire distribuée uniformément sur l'intervalle $(2, 4)$ pour chaque distributeur d'essence, indépendamment d'un automobiliste à l'autre (et que les divers temps de service sont indépendants des temps entre les arrivées des automobilistes). Quelles sont les probabilités limites du système? Quelle est la variance du nombre de voitures dans la station-service en équilibre, si $\lambda = 1/3$?

Solution. On a un système de file d'attente $M/G/2/2$ pour lequel $S \sim \mathrm{U}(2,4)$, de sorte que $E[S] = 3$. En utilisant l'équation (9.26), on peut écrire que

$$\pi_0 = \left(\sum_{k=0}^{2} \frac{(3\lambda)^k}{k!}\right)^{-1} = \left(1 + 3\lambda + \frac{(3\lambda)^2}{2}\right)^{-1}$$

et

$$\pi_1 = 3\lambda\pi_0 \quad \text{et} \quad \pi_2 = \frac{(3\lambda)^2}{2}\pi_0$$

Si $\lambda = 1/3$, on a:

$$\pi_0 = \left(1 + 1 + \frac{1}{2}\right)^{-1} = \frac{2}{5}, \quad \pi_1 = \pi_0 = \frac{2}{5} \quad \text{et} \quad \pi_2 = \frac{1}{2}\pi_0 = \frac{1}{5}$$

Il s'ensuit que

$$E[N] = 1 \times \pi_1 + 2 \times \pi_2 = \frac{4}{5} \quad \text{et} \quad E[N^2] = 1 \times \pi_1 + 4 \times \pi_2 = \frac{6}{5}$$

de sorte que

$$\mathrm{VAR}[N] = \frac{6}{5} - \left(\frac{4}{5}\right)^2 = \frac{14}{25}$$

Exercices

Question n° 1

Soit $\{N_i(t), t \geq 0\}$ un processus de Poisson de taux λ_i, pour $i = 1, 2$. On suppose que les deux processus de Poisson sont indépendants. On définit

$$X(t) = N_1(t) - N_2(t) \quad \text{pour } t \geq 0$$

Le processus stochastique $\{X(t), t \geq 0\}$ est-il une chaîne de Markov à temps continu? Est-ce un processus de naissance et de mort (avec espace des états $\mathbb{Z} := \{\ldots, -2, 1, 0, 1, 2, \ldots\}$)? Justifier les réponses.

Question n° 2

Soit $\{N(t), t \geq 0\}$ un processus de comptage (voir l'exemple 9.1.1) pour lequel le temps T jusqu'au premier événement, et entre deux événements consécutifs, présente une distribution uniforme sur l'intervalle $(0, 1)$. Montrer que le processus stochastique $\{N^*(t), t \geq 0\}$, où $N^*(0) := 0$ et

$$N^*(t) := N(-\ln t) \quad \text{pour } t > 0$$

est un processus de naissance pur.

Indication. Voir l'exemple 3.4.4.

Question n° 3

On considère le processus de naissance et de mort $\{X(t), t \geq 0\}$ ayant les taux de naissance et de mort $\lambda_n = \lambda$ et $\mu_n = n\mu$, pour $n = 0, 1, 2, \ldots$ Calculer, si elles existent, les probabilités limites du processus.

Question n° 4

Soit $\{N(t), t \geq 0\}$ un processus de Poisson de taux $\lambda = \ln 2$. On définit $W_i = \text{ent}(\tau_i + 1)$, pour $i = 0, 1, \ldots$, où τ_i est le temps que le processus passe dans l'état i, et *ent* désigne la *partie entière*. On peut montrer que W_i présente une distribution géométrique de paramètre $p = 1 - e^{-\lambda}$. Calculer les probabilités (a) $P[W_0 = W_1]$; (b) $P[W_0 > W_1]$; (c) $P[W_0 > W_1 \mid W_1 > 1]$.

Question n° 5

On suppose que $\{X(t), t \geq 0\}$ est un processus de Yule avec $\lambda = 2$ (voir l'exercice résolu n° 2). Calculer (a) $E[\tau_1^4 + \tau_2^4 + \tau_3^4]$; (b) $E[\tau_1 + \tau_2^2 + \tau_3^3]$; (c) le coefficient de corrélation de τ_1^2 et τ_1^3.

Indication. On a (voir l'exemple 3.5.2) que

$$\int_0^\infty x^n \lambda e^{-\lambda x} dx = \frac{n!}{\lambda^n} \quad \text{pour } n = 0, 1, \ldots$$

Question n° 6

Des clients arrivent à un certain magasin selon un processus de Poisson de taux $\lambda = 1/2$ par minute. On suppose que chaque client demeure exactement cinq minutes dans le magasin.

(a) Quelle est l'espérance mathématique du nombre de clients dans le magasin à l'instant $t = 60$?

(b) Quelle est l'espérance mathématique du nombre de clients en (a), étant donné que le magasin n'est pas vide à l'instant $t = 60$?

(c) Quelle est l'espérance mathématique du nombre de clients en (a), étant donné que le nombre de clients dans le magasin à l'instant $t = 60$ n'est pas égal à 1?

Question n° 7

Quel est le nombre moyen de clients dans un système de file d'attente $M/M/1$ en équilibre, étant donné que le nombre de clients est un nombre impair?

Question n° 8

On considère un modèle d'attente $M/M/1$ en équilibre, avec $\lambda = \mu/2$. Quelle est la probabilité qu'il y ait plus de cinq clients dans le système, étant donné qu'il y en a au moins deux?

Question n° 9

On suppose que la politique de service pour un modèle de file d'attente $M/M/1$ est la suivante: lorsque le serveur finit de servir un client, le prochain à être servi est pris au hasard parmi ceux qui font la queue. Quelle est l'espérance

mathématique du temps total qu'un client qui est arrivé alors que le système (en équilibre) était dans l'état 3 a passé dans le système, étant donné qu'aucun client qui est (possiblement) arrivé après le client en question n'a été servi avant lui?

Question n^{o} 10

Calculer la variance du temps total T qu'un client quelconque passe dans un système de file d'attente $M/M/1$ en équilibre, étant donné que $1 < T < 2$, si $\lambda = 1$ et $\mu = 2$.

Question n^{o} 11

On suppose qu'après avoir été servi, chaque client dans un système de file d'attente $M/M/1/2$ retourne immédiatement (exactement une fois) devant le serveur (s'il y a un client qui fait la queue, alors le client qui retourne doit attendre que le serveur soit libre). Définir un espace des états approprié et écrire les équations d'équilibre du système.

Question n^{o} 12

Un client qui n'a pas pu entrer dans un système de file d'attente $M/M/1/c$ (en équilibre) à l'instant t_0 décide de revenir à l'instant $t_0 + 2$. Quelle est la probabilité que le client en question puisse alors entrer dans le système, étant donné qu'exactement un client est arrivé dans l'intervalle $(t_0, t_0 + 2)$ et n'a pas pu entrer dans le système également?

Question n^{o} 13

On suppose que pour un système de file d'attente qui, sans cette modification, serait un modèle $M/M/1/2$, un client qui entre dans le système, mais doit attendre avant d'être servi, décide de quitter s'il attend encore après un temps aléatoire présentant une distribution exponentielle de paramètre θ. De plus, ce temps aléatoire est indépendant du temps de service et des temps entre les arrivées. Soit $X(t)$ le nombre de clients dans le système à l'instant t. Le processus stochastique $\{X(t), t \geq 0\}$ est une chaîne de Markov à temps continu (c'est un processus de naissance et de mort, plus précisément). (a) Écrire les équations d'équilibre du processus. (b) Calculer les probabilités limites dans le cas où $\lambda = \mu = \theta = 1$.

Question n^{o} 14

On suppose que la probabilité que le serveur, dans un système de file d'attente $M/M/1/3$, soit incapable de fournir le service demandé par un client quelconque est égale à $p \in (0, 1)$, indépendamment d'un client à l'autre. On suppose aussi que le temps que met le serveur à décider s'il sera capable de servir de façon

satisfaisante ou non un client donné est une variable aléatoire qui présente une distribution exponentielle de paramètre μ_0. Définir un espace des états approprié et écrire les équations d'équilibre du système.

Question n° 15

On considère un système de file d'attente $M/M/2$. On suppose que n'importe quel client qui accepte de payer deux fois plus qu'un client ordinaire pour son service sera servi au taux 2μ. Si un client de ce type arrive alors qu'il y a exactement deux clients déjà présents dans le système, quelle est la probabilité que ce nouveau client passe moins de $1/\mu$ unité(s) de temps (c'est-à-dire le temps moyen de service d'un client ordinaire) dans le système?

Question n° 16

On suppose que lorsqu'un système de file d'attente $M/M/2$ est vide, un client qui arrive se dirige vers le serveur n° 1 avec une probabilité de 1/2. Quelle est la probabilité que les deux premiers clients, à partir de l'instant initial, soient servis par (a) le serveur n° 1? (b) des serveurs différents?

Question n° 17

Soit $X(t)$ le nombre de clients dans un système de file d'attente $M/M/2/4$. On suppose que t_0 est assez grand pour que le système soit en régime stationnaire. Calculer l'espérance conditionnelle $E[X(t_0) \mid X(t_0) > 0]$, si $\lambda = \mu/2$.

Question n° 18

Pour un modèle d'attente $M/M/4/4$ en équilibre, avec $\lambda = \mu$, quelle est la variance du nombre de clients dans le système, étant donné qu'il n'est pas vide?

Question n° 19

On suppose que dans un certain système de file d'attente $M/M/2/3$, le temps de service d'un client quelconque est une variable aléatoire qui présente une distribution exponentielle de paramètre μ. Cependant, si un serveur n'a pas terminé son service après un temps aléatoire (indépendant du temps de service réel) présentant une distribution exponentielle de paramètre μ_0, alors le client doit quitter le système. (a) Écrire les équations d'équilibre du système. (b) Quelle est la proportion des clients qui doivent quitter le système avant d'avoir été servis complètement?

Question n° 20

Pour un modèle d'attente $M/M/5/5$ avec $\lambda = \mu$, quelle est l'espérance mathématique de $1/N$, où N est le nombre de clients dans le système en régime stationnaire, étant donné que le système n'est ni vide ni plein?

Questions à choix multiple

Question n° 1

Soit $\{X(t), t \geq 0\}$ un processus de naissance et de mort dont les taux sont $\lambda_n \equiv 1$ et $\mu_n \equiv 2$. On suppose que $X(0) = 0$. Quelle est la probabilité que le processus retourne exactement deux fois à l'état 0 avant de visiter l'état 2?

(a) 1/27 (b) 4/27 (c) 2/9 (d) 1/3 (e) 4/9

Question n° 2

On considère le processus de mort pur $\{X(t), t \geq 0\}$ dont les taux sont $\mu_n \equiv 1$. Calculer la probabilité $P[X(5) = 0 \mid X(0) = 5]$.

(a) 0,1755 (b) 0,3840 (c) 0,4405 (d) 0,5595 (e) 0,6160

Question n° 3

La chaîne de Markov à temps continu $\{X(t), t \geq 0\}$, dont l'espace des états est $\{0, 1, 2\}$, est telle que $\nu_i \equiv \nu$, $\rho_{0,1} = 1/2$, $\rho_{1,0} = 1/4$ et $\rho_{2,0} = 1/4$. Calculer la probabilité limite que le processus ne soit pas dans l'état 0.

(a) 1/5 (b) 2/5 (c) 1/2 (d) 3/5 (e) 4/5

Question n° 4

Soit T le temps total que passe un client quelconque dans un modèle d'attente $M/M/1$ avec $\lambda = 1$ et $\mu = 3$. Calculer $E[T^2 \mid T > t_0]$, où $t_0 > 0$.

(a) $\frac{1}{4} + t_0 + t_0^2$ (b) $\frac{1}{2} + t_0 + t_0^2$ (c) $1 + t_0 + t_0^2$ (d) 1/4 (e) 1/2

Question n° 5

Quelle est la probabilité qu'un client, qui a quitté un modèle d'attente $M/M/1$ avec $\lambda = 1$ et $\mu = 2$ avant l'arrivée du client suivant, ait passé moins d'une unité de temps dans le système?

(a) $\frac{1}{2}\left(1 - e^{-1}\right)$ (b) $\frac{1}{2}\left(1 - e^{-2}\right)$ (c) $1 - e^{-1}$ (d) $1 - e^{-2}$ (e) $1 - \frac{1}{2}e^{-2}$

Question n° 6

Pour un système de file d'attente $M/M/1/2$, quelle est la probabilité que le système ait été plein avant que le premier départ ait eu lieu, étant donné que le second client est arrivé à l'instant $t = 2$?

(a) $\frac{1}{2\mu}\left(1 - e^{-2\mu}\right)$ (b) $\frac{1}{\mu}\left(1 - e^{-2\mu}\right)$ (c) $\frac{1}{2\mu}\left(1 - e^{-\mu}\right)$ (d) $\frac{1}{\mu}\left(1 - e^{-\mu}\right)$
(e) $\frac{1}{\mu}\left(1 - e^{-\mu} - e^{-2\mu}\right)$

Question n° 7

On suppose que le système de file d'attente $M/M/1/2$ est modifié comme suit: chaque fois que le serveur finit de servir un client, il est indisponible pour

un temps aléatoire τ (indépendant du temps de service et des temps entre les arrivées) qui présente une distribution exponentielle de paramètre θ. Calculer la probabilité limite que le serveur soit occupé à servir un client, si $\lambda = \mu = 1$ et $\theta = 1/2$.

Indication. Définir les états suivants:

0: le système est vide; le serveur est disponible
0^*: le système est vide; le serveur est indisponible
1: un client est en train d'être servi; personne ne fait la queue
1^*: un client attend d'être servi; le serveur est indisponible
2: un client est en train d'être servi et un autre fait la queue
2^*: deux clients font la queue; le serveur est indisponible

(a) 3/49 (b) 13/49 (c) 16/49 (d) 19/49 (e) 20/49

Question n° 8

On suppose qu'il y a cinq clients dans un système de file d'attente $M/M/2$ à un instant donné. Trouver la probabilité que les trois clients qui font la queue ne soient pas servis par le même serveur.

(a) 3/8 (b) 1/2 (c) 5/8 (d) 3/4 (e) 7/8

Question n° 9

Quel est le nombre moyen de clients dans un système de file d'attente $M/M/2/4$ en équilibre, avec $\lambda = \mu$, étant donné qu'il n'est ni plein ni vide?

(a) 9/7 (b) 11/7 (c) 2 (d) 15/7 (e) 17/7

Question n° 10

On suppose que le temps de service pour un système de file d'attente $M/G/3/3$, avec $\lambda = 1$, est une variable aléatoire S définie par $S = Z^4$, où $Z \sim \mathrm{N}(0,1)$. Quel est le nombre moyen de clients dans le système en régime stationnaire?

Indication. Le carré d'une variable aléatoire $\mathrm{N}(0,1)$ présente une distribution gamma de paramètres $\alpha = \lambda = 1/2$.

(a) 18/13 (b) 41/26 (c) 51/26 (d) 2 (e) 2,5

A

Tableaux statistiques

Tableau A.1. Fonction de répartition de la distribution binomiale

		p					
		0,05	0,10	0,20	0,25	0,40	0,50
n	x						
2	0	0,9025	0,8100	0,6400	0,5625	0,3600	0,2500
	1	0,9975	0,9900	0,9600	0,9375	0,8400	0,7500
3	0	0,8574	0,7290	0,5120	0,4219	0,2160	0,1250
	1	0,9927	0,9720	0,8960	0,8438	0,6480	0,5000
	2	0,9999	0,9990	0,9920	0,9844	0,9360	0,8750
4	0	0,8145	0,6561	0,4096	0,3164	0,1296	0,0625
	1	0,9860	0,9477	0,8192	0,7383	0,4752	0,3125
	2	0,9995	0,9963	0,9728	0,9493	0,8208	0,6875
	3	1,0000	0,9999	0,9984	0,9961	0,9744	0,9375
5	0	0,7738	0,5905	0,3277	0,2373	0,0778	0,0313
	1	0,9774	0,9185	0,7373	0,6328	0,3370	0,1875
	2	0,9988	0,9914	0,9421	0,8965	0,6826	0,5000
	3	1,0000	0,9995	0,9933	0,9844	0,9130	0,8125
	4	1,0000	1,0000	0,9997	0,9990	0,9898	0,9688
10	0	0,5987	0,3487	0,1074	0,0563	0,0060	0,0010
	1	0,9139	0,7361	0,3758	0,2440	0,0464	0,0107
	2	0,9885	0,9298	0,6778	0,5256	0,1673	0,0547
	3	0,9990	0,9872	0,8791	0,7759	0,3823	0,1719
	4	0,9999	0,9984	0,9672	0,9219	0,6331	0,3770
	5	1,0000	0,9999	0,9936	0,9803	0,8338	0,6230
	6		1,0000	0,9991	0,9965	0,9452	0,8281
	7			0,9999	0,9996	0,9877	0,9453
	8			1,0000	1,0000	0,9983	0,9893
	9					0,9999	0,9990
15	0	0,4633	0,2059	0,0352	0,0134	0,0005	0,0000
	1	0,8290	0,5490	0,1671	0,0802	0,0052	0,0005
	2	0,9638	0,8159	0,3980	0,2361	0,0271	0,0037
	3	0,9945	0,9444	0,6482	0,4613	0,0905	0,0176
	4	0,9994	0,9873	0,8358	0,6865	0,2173	0,0592
	5	0,9999	0,9977	0,9389	0,8516	0,4032	0,1509
	6	1,0000	0,9997	0,9819	0,9434	0,6098	0,3036
	7		1,0000	0,9958	0,9827	0,7869	0,5000
	8			0,9992	0,9958	0,9050	0,6964
	9			0,9999	0,9992	0,9662	0,8491
	10			1,0000	0,9999	0,9907	0,9408
	11				1,0000	0,9981	0,9824
	12					0,9997	0,9963
	13					1,0000	0,9995
	14						1,0000
20	0	0,3585	0,1216	0,0115	0,0032	0,0000	
	1	0,7358	0,3917	0,0692	0,0243	0,0005	0,0000
	2	0,9245	0,6769	0,2061	0,0913	0,0036	0,0002
	3	0,9841	0,8670	0,4114	0,2252	0,0160	0,0013
	4	0,9974	0,9568	0,6296	0,4148	0,0510	0,0059
	5	0,9997	0,9887	0,8042	0,6172	0,1256	0,0207
	6	1,0000	0,9976	0,9133	0,7858	0,2500	0,0577
	7		0,9996	0,9679	0,8982	0,4159	0,1316
	8		0,9999	0,9900	0,9591	0,5956	0,2517
	9		1,0000	0,9974	0,9861	0,7553	0,4119
	10			0,9994	0,9961	0,8725	0,5881
	11			0,9999	0,9991	0,9435	0,7483
	12			1,0000	0,9998	0,9790	0,8684
	13				1,0000	0,9935	0,9423
	14					0,9984	0,9793
	15					0,9997	0,9941
	16					1,0000	0,9987
	17						0,9998
	18						1,0000

Tableau A.2. Fonction de répartition de la distribution de Poisson

	λ							
	0,5	1	1,5	2	5	10	15	20
x								
0	0,6065	0,3679	0,2231	0,1353	0,0067	0,0000		
1	0,9098	0,7358	0,5578	0,4060	0,0404	0,0005		
2	0,9856	0,9197	0,8088	0,6767	0,1247	0,0028	0,0000	
3	0,9982	0,9810	0,9344	0,8571	0,2650	0,0103	0,0002	
4	0,9998	0,9963	0,9814	0,9473	0,4405	0,0293	0,0009	0,0000
5	1,0000	0,9994	0,9955	0,9834	0,6160	0,0671	0,0028	0,0001
6		0,9999	0,9991	0,9955	0,7622	0,1301	0,0076	0,0003
7		1,0000	0,9998	0,9989	0,8666	0,2202	0,0180	0,0008
8			1,0000	0,9998	0,9319	0,3328	0,0374	0,0021
9				1,0000	0,9682	0,4579	0,0699	0,0050
10					0,9863	0,5830	0,1185	0,0108
11					0,9945	0,6968	0,1848	0,0214
12					0,9980	0,7916	0,2676	0,0390
13					0,9993	0,8645	0,3632	0,0661
14					0,9998	0,9165	0,4657	0,1049
15					0,9999	0,9513	0,5681	0,1565
16					1,0000	0,9730	0,6641	0,2211
17						0,9857	0,7489	0,2970
18						0,9928	0,8195	0,3814
19						0,9965	0,8752	0,4703
20						0,9984	0,9170	0,5591
21						0,9993	0,9469	0,6437
22						0,9997	0,9673	0,7206
23						0,9999	0,9805	0,7875
24						1,0000	0,9888	0,8432
25							0,9938	0,8878
26							0,9967	0,9221
27							0,9983	0,9475
28							0,9991	0,9657
29							0,9996	0,9782
30							0,9998	0,9865
31							0,9999	0,9919
32							1,0000	0,9953

Tableau A.3. Valeurs de la fonction $\Phi(z)$

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
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,7	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,8	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999
3,9	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Tableau A.4. Valeurs de la fonction $Q^{-1}(p)$ pour quelques valeurs de p

p	0,10	0,05	0,01	0,005	0,001	0,0001	0,00001
$Q^{-1}(p)$	1,282	1,645	2,326	2,576	3,090	3,719	4,265

B

Quantiles des distributions d'échantillonnage

Distribution normale

α	0,25	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
z_α	0,674	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Distribution de Student

n	1	2	3	4	5	6	7	8
$t_{0,005;n}$	63,657	9,925	5,841	4,604	4,032	3,707	3,499	3,355
$t_{0,01;n}$	31,821	6,965	4,541	3,747	3,365	3,143	2,998	2,896
$t_{0,025;n}$	12,706	4,303	3,182	2,776	2,571	2,447	2,365	2,306
$t_{0,05;n}$	6,314	2,920	2,353	2,132	2,015	1,943	1,895	1,860
$t_{0,10;n}$	3,078	1,886	1,638	1,533	1,476	1,440	1,415	1,397

n	9	10	15	20	25	30	40	∞
$t_{0,005;n}$	3,250	3,169	2,947	2,845	2,787	2,750	2,704	2,576
$t_{0,01;n}$	2,821	2,764	2,602	2,528	2,485	2,457	2,423	2,326
$t_{0,025;n}$	2,262	2,228	2,131	2,086	2,060	2,042	2,021	1,960
$t_{0,05;n}$	1,833	1,812	1,753	1,725	1,708	1,697	1,684	1,645
$t_{0,10;n}$	1,383	1,372	1,341	1,325	1,316	1,310	1,303	1,282

Distribution du khi-deux

n	1	2	3	4	5	6	7	8	9
$\chi^2_{0,005;n}$	7,88	10,60	12,84	14,86	16,75	18,55	20,28	21,96	23,59
$\chi^2_{0,01;n}$	6,63	9,21	11,34	13,28	15,09	16,81	18,48	20,09	21,67
$\chi^2_{0,025;n}$	5,02	7,38	9,35	11,14	12,83	14,45	16,01	17,53	19,02
$\chi^2_{0,05;n}$	3,84	5,99	7,81	9,49	11,07	12,59	14,07	15,51	16,92
$\chi^2_{0,10;n}$	2,71	4,61	6,25	7,78	9,24	10,65	12,02	13,36	14,68
$\chi^2_{0,90;n}$	0,02	0,21	0,58	1,06	1,61	2,20	2,83	3,49	4,17
$\chi^2_{0,95;n}$	0^+	0,10	0,35	0,71	1,15	1,64	2,17	2,73	3,33
$\chi^2_{0,975;n}$	0^+	0,05	0,22	0,48	0,83	1,24	1,69	2,18	2,70
$\chi^2_{0,99;n}$	0^+	0,02	0,11	0,30	0,55	0,87	1,24	1,65	2,09
$\chi^2_{0,995;n}$	0^+	0,01	0,07	0,21	0,41	0,68	0,99	1,34	1,73

n	10	15	20	25	30	40	50	100
$\chi^2_{0,005;n}$	25,19	32,80	40,00	46,93	53,67	66,77	79,49	140,17
$\chi^2_{0,01;n}$	23,21	30,58	37,57	44,31	50,89	63,69	76,15	135,81
$\chi^2_{0,025;n}$	20,48	27,49	34,17	40,65	46,98	59,34	71,42	129,56
$\chi^2_{0,05;n}$	18,31	25,00	31,41	37,65	43,77	55,76	67,50	124,34
$\chi^2_{0,10;n}$	15,99	22,31	28,41	34,28	40,26	51,81	63,17	118,50
$\chi^2_{0,90;n}$	4,87	8,55	12,44	16,47	20,60	29,05	37,69	82,36
$\chi^2_{0,95;n}$	3,94	7,26	10,85	14,61	18,49	26,51	34,76	77,93
$\chi^2_{0,975;n}$	3,25	6,27	9,59	13,12	16,79	24,43	32,36	74,22
$\chi^2_{0,99;n}$	2,56	5,23	8,26	11,52	14,95	22,16	29,71	70,06
$\chi^2_{0,995;n}$	2,16	4,60	7,43	10,52	13,79	20,71	27,99	67,33

Distribution de Fisher

n	1	2	3	4	5	6	7	8
$F_{0,01;n,n}$	4052	99,00	29,46	15,98	10,97	8,47	6,99	6,03
$F_{0,025;n,n}$	647,8	39,00	15,44	9,60	7,15	5,82	4,99	4,43
$F_{0,05;n,n}$	161,4	19,00	9,28	6,39	5,05	4,28	3,79	3,44
$F_{0,10;n,n}$	39,86	9,00	5,39	4,11	3,45	3,05	2,78	2,59

n	9	10	15	20	30	40	50	100	200
$F_{0,01;n,n}$	5,35	4,85	3,52	2,94	2,39	2,11	1,95	1,60	1,39
$F_{0,025;n,n}$	4,03	3,72	2,86	2,46	2,07	1,88	1,75	1,48	1,32
$F_{0,05;n,n}$	3,18	2,98	2,40	2,12	1,84	1,69	1,60	1,39	1,26
$F_{0,10;n,n}$	2,44	2,32	1,97	1,79	1,61	1,51	1,44	1,29	1,20

Valeurs de $F_{0,025;n_1,n_2}$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	11	12
1	648	800	864	900	922	937	948	957	963	969	973	977
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,4
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,5	14,4	14,4	14,3
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,79	8,75
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,57	6,52
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,41	5,37
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,71	4,67
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,24	4,20
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,91	3,87
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,66	3,62
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,47	3,43
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,32	3,28

Valeurs de $F_{0,05;n_1,n_2}$

$n_2 \backslash n_1$	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4
3	10,1	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

C

Réponses des exercices à numéros pairs

Chapitre 1

2. $F(x)$ est continue en n'importe quel $x \in \mathbb{R}$, sauf en $x = \{0, 1, 2\}$. En ces points, elle est seulement continue à droite.
4. Discontinue.
6. xe^{-x}.
8. $f'(x) = \frac{2}{3}x^{-1/3}$; $x = 0$.
10. $\sqrt{2\pi}$.
12. $-\frac{1}{2}e^{-x}(\cos x + \sin x)$.
14. 3/8.
16. −
18. $\ln 2$.
20. (a) $\frac{p}{1-(1-p)z}$; (b) $k!(1-p)^k p$.

Chapitre 2

2. (a) 7/12; (b) 55/72; (c) 2/9; (d) 3/8.
4. (a) $P[\{\omega_1\}] = 0{,}1$; $P[\{\omega_2\}] = 0{,}45$; $P[\{\omega_3\}] = 0{,}0171$; $P[\{\omega_4\}] = 0{,}8379$;
(b) (i) $R = A_1 \cup (A_1^c \cap A_2) \cup (A_1^c \cap A_2^c \cap A_3)$; (ii) 0,1621; (iii) 0,2776;
(c) (i) $B = R_1^c \cup R_2^c \cup R_3^c$; (ii) 0,9957.
6. 4/5.
8. 0,9963.
10. (a) 1/10; (b) 1/10.
12. (a) 0,325; (b) 0,05; (c) 0,25; (d) 0,66375.

14. 48/95.
16. 23/32.
18. 9.
20. 63,64 %.
22. (a) 0,8829; (b) 0,9083.
24. (a) 0,9972; (b) (i) 0,8145; (ii) 0,13.
26. (a) $-$; (b) $\frac{1}{n-1}$; (c) 0,8585.
28. (a) 1/64; (b) 3/32.
30. (a) 86.400; (b) 3840.

Chapitre 3

2. (a) Hyp($N = 100$, $n = 2$, $d =$ nombre d'articles défectueux dans le lot); (b) 0,9602; (c) 0,9604; (d) 0,9608.
4. (a) Poi(3); 0,616; (b) Poi(9); 0,979; (c) Exp(3); $1 - e^{-3}$; (d) Exp(3); e^{-3}; (e) G($\alpha = 2, \lambda = 3$); (f) B($n = 100, p = 0{,}05$); 0,037.
6. (a) (i) B($20, \theta$); (ii) Hyp($1000, 20, 1000\,\theta$); (b) (i) 0,010; (ii) 0,032.
8. (a) $X \sim$ Poi(1,2); $Y \sim$ Géom(0,301); $Z \sim$ B(12; 0,301); (b) 4^{e}; (c) 0,166; (d) 0,218.
10. 0,7746.
12. 0,6233.
14. 0,1215.
16. 0,759.
18. 0,4.
20. $1 - e^{-5}$.
22. (a)

$$f_X(x) = \begin{cases} 1/2 \text{ si } 0 \leq x \leq 1 \\ 1/6 \text{ si } 1 < x < 4 \\ 0 \quad \text{ailleurs} \end{cases}$$

(b) 2,5; (c) 1,5; (d) ∞; (e) (i) 1/2; (ii) 1.
24. (a) 0,657; (b) 0,663; (c) 7/17.
26. (a) 0,3754; (b) 0,3679.
28. (a) 3; (b)

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 3x^2 - 2x^3 & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x > 1 \end{cases}$$

(c) $\left(\frac{27}{32}\right)2^k$ pour $k = 1, 2, \ldots$; (d)

$$f_Z(z) = \begin{cases} 3(1-\sqrt{z}) \text{ si } 0 \leq z \leq 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

30. (a) 6; (b) 7; (c) 0,8.
32. 5/8.
34. 0,00369.
36. 0,027.
38. $1 - e^{-\lambda y}$ si $y = 1, 2, \ldots$
40. 2,963.
42. Deux moteurs: fiabilité $\simeq$ 0,84; quatre moteurs: fiabilité $\simeq$ 0,821. Donc, fiabilité supérieure ici avec deux moteurs.
44. (a) 0,75; (b) 0,2519.
46. (a) (i) 0,7833; (ii) 0,7558; (b) 0,7361.
48. (a) $E[X] = (\pi k/2)^{1/2}$; $\text{VAR}[X] = 2k\left(1 - \frac{\pi}{4}\right)$; (b) k est un paramètre d'échelle.
50. (a) $\sqrt{\pi}/8$; (b) $E\left[e^{X^2/2}\right] = 1{,}25$; $\text{VAR}\left[e^{X^2/2}\right] = \infty$.
52. (a)

$$B(W) = \begin{cases} v - \alpha - \beta W \text{ si } W \geq w_0 \\ c - \alpha - \beta W \text{ si } W < w_0 \end{cases}$$

(b) $w_0 + \left[-2\sigma^2 \ln\left(\dfrac{\beta\sigma\sqrt{2\pi}}{v-c}\right)\right]^{1/2}$.

Chapitre 4

2. (a) 3; (b)

$$f_X(x) = \begin{cases} 3x^2 \text{ si } 0 < x < 1 \\ 0 \quad \text{ailleurs} \end{cases} \qquad f_Y(y) = \begin{cases} \frac{3}{2}(1-y^2) \text{ si } 0 < y < 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

(c) $\text{VAR}[X] = 3/80$; $\text{VAR}[Y] = 19/320$; (d) 0,3974.
4. (a) $\text{N}(750\,\theta; 187{,}5\,\theta^2)$; (b) 220 \$.
6. (a) (i)

$$f_X(x) = \begin{cases} \frac{3}{4}(1-x^2) \text{ si } -1 \leq x \leq 1 \\ 0 \qquad \text{ailleurs} \end{cases} \qquad f_Y(y) = \begin{cases} \frac{3}{2}\sqrt{y} \text{ si } 0 \leq y < 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

(ii) 0; (b) non, car $f_X(x)f_Y(y)$ n'est pas identique à 3/4 ($\equiv f_{X,Y}(x,y)$).
8. 1,5.
10. 1/64.
12. 0,1586.
14. $-1/11$.
16. 0,8665.
18. (a)

$$f_X(x) = \begin{cases} 6x(1-x) \text{ si } 0 < x < 1 \\ 0 \qquad \text{ailleurs} \end{cases} \qquad f_Y(y) = \begin{cases} 3y^2 \text{ si } 0 < y < 1 \\ 0 \quad \text{ailleurs} \end{cases}$$

(b) 5/16.
20. (a) 0,5987; (b) 1,608; (c) 0,0871.
22. (a) 0,6065; (b) 7,29; (c) 0,1428; (d) 0,3307; (e) 0,7291.
24. (a) 0,2048; (b) 3/5; (c) $e^{-3/5}$; (d) 0,0907.
26. (a) 0,6587; (b) 3,564.
28. (a) 0,157; (b) 0,3413; $X = \sum_{i=1}^{25} X_i$, où $X_i \sim \text{Exp}(1/2)$ pour tout i et les X_i sont des variables aléatoires indépendantes.
30. 10/3.
32. 18.
34. (a) $-$; (b)

$$f_X(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-x^2} \text{ si } -1 \leq x \leq 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

et $f_Y(y) = f_X(y)$, par symétrie; (c) non, car $f_X(x)f_Y(y)$ n'est pas identique à $1/\pi$ ($\equiv f_{X,Y}(x,y)$); (d) 3/4.
36. (a) 0,045; (b) 527,08.
38. (a) $T \sim \text{Exp}(1/5)$; (b)

$$f_T(t) = \begin{cases} \frac{1}{5}(1-e^{-t/50})^9 e^{-t/50} \text{ si } t \geq 0 \\ 0 \qquad\qquad\qquad \text{ailleurs} \end{cases}$$

(c) $T \sim \text{G}(\alpha = 10, \lambda = 1/50)$; c'est-à-dire que

$$f_T(t) = \begin{cases} \dfrac{1}{50^{10}\Gamma(10)} t^9 e^{-t/50} \text{ si } t \geq 0 \\ 0 \qquad\qquad\qquad\qquad\quad \text{ailleurs} \end{cases}$$

40. (a)

$$f_Y(y) = \begin{cases} 3(y-1)^2 \text{ si } 0 < y < 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

(b) 0,1566; (c)

$$f_Z(z) = \begin{cases} (1 - z^{-1/3})^2 \text{ si } 0 < z < 1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

d) 1/2.

42. (a) (i)

x_2	0	1
$p_{X_2}(x_2)$	7/12	5/12

(ii) 1/3 si $x_2 = 0$ et 2/3 si $x_2 = 1$; (b)

y	-1	0	3	8
$F_Y(y)$	1/6	1/2	5/6	1

44. (a) 0,95; (b) (i)

$$P[Y \leq y \mid X_1 = x_1] = \begin{cases} 0 & \text{si } y \leq x_1 \\ 1 - e^{-(y-x_1)/2} & \text{si } y > x_1 \end{cases}$$

(ii)

$$f_Y(y \mid X_1 = x_1) = \begin{cases} \frac{1}{2} e^{-(y-x_1)/2} \text{ si } y > x_1 \\ 0 \qquad \text{ailleurs} \end{cases}$$

46. (a) $X \sim \text{Poi}(1)$; $Y \sim \text{Poi}(1/2)$; $Y \mid \{X = 35\} \sim \text{B}(n = 35, p = \frac{1}{2})$;
(b) (i) 0,5873; (ii) 0,5876; (c) 0,6970.

48. (a)

$$f_X(x) = \begin{cases} \dfrac{4-x}{8} \text{ si } 0 \leq x \leq 4 \\ 0 \quad \text{ailleurs} \end{cases}$$

(b) (i) 4/3; (ii) 8/9; (iii) 0,32; (iv) 2,4; (c) $-1/2$; X et Y ne sont donc pas indépendantes, puisque $\rho_{X,Y} \neq 0$.

50. (a)

$$f_X(x) = \begin{cases} 15x^2 - 45x^4 + 30x^5 \text{ si } 0 < x < 1 \\ 0 \qquad\qquad \text{ailleurs} \end{cases}$$

$$f_Y(y) = \begin{cases} 30y^4(1-y) \text{ si } 0 < y < 1 \\ 0 \qquad\qquad \text{ailleurs} \end{cases}$$

(b) non, car, par exemple, $f_{X,Y}\left(1, \frac{1}{2}\right) = 22{,}5 \neq f_X(1) f_Y\left(\frac{1}{2}\right) = 0$; (c) 0,0191.

52. (a)

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{2(x+1)} & \text{si } -1 \leq x \leq 1, -1 \leq y \leq x \\ 0 & \text{ailleurs} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{2}[\ln 2 - \ln(y+1)] & \text{si } -1 \leq y \leq 1 \\ 0 & \text{ailleurs} \end{cases}$$

(b) (i) $-1/3$; (ii) $-1/2$; (c) $1/6$.

54. (a)

$x_2 \backslash x_1$	-1	0	1
-1	1/64	6/64	1/64
0	6/64	36/64	6/64
1	1/64	6/64	1/64

(b) $-\sqrt{2}/2$; (c) non, car $\rho_{X,Y} \neq 0$.

Chapitre 5

2. (a) χ^2_1; (b) 3,841; (c) t_1; (d) $-31{,}821$.

4. (a) $n_{A_1} = 22$, $n_{A_2} = 26$, $n_{A_3} = 12$, $n_{B_1} = 27$, $n_{B_2} = 27$, $n_{B_3} = 36$;
(b) (i) $\bar{d}_A \simeq 3{,}67 \times 10^3$ km et $\bar{d}_B = 4{,}2 \times 10^3$ km; donc B; (ii) $CV_A \simeq 40{,}32$ % et $CV_B \simeq 39{,}77$ %; donc B.

6. 26.

8. $-0{,}9839$.

10. 0.

12. (a) 3,03; (b) 2,87 (avec x_{11} égal à la moyenne des 10 premières observations);
(c) $\tilde{x}_{\min} = 5$ et $\tilde{x}_{\max} = 6$.

14. 1,33.

16. (a) 1,06; (b) 21,666; (c) 4,15.

18. (a) 0,0116 (en utilisant une interpolation linéaire); (b) 0,0112.

20. 0,2810.

22. (a) 4; (b) 25; (c) 3.

24. (a) $\bar{x} = 20{,}65$; $s^2 \simeq 22{,}34$; $\hat{\beta}_1 \simeq 0{,}337$; $\hat{\beta}_2 \simeq 1{,}93$; (b) 17; (c) Poi($\lambda \simeq 20$).

26. (a) $\bar{x} = 1{,}47$; $s^2 \simeq 0{,}9587$; $\hat{\beta}_1 \simeq 0{,}0326$; $\hat{\beta}_2 \simeq 3{,}3273$; (b) (i) 0,9; (ii) 0,8424; (iii) 139.

28. (a) 0,3173; (b) 0,0202; (c) 0,0105; (d) 0,9992.

30. (a) 0,9829; (b) 0,97725.

32. $\dfrac{n}{2\sum_{i=1}^{n}|X_i|}$.

34. (a) $-$; (b) $\hat{p}_2$, car E.Q.M.$[\hat{p}_1] = p(1-p)$ et E.Q.M.$[\hat{p}_2] = \frac{1}{2}p(1-p)$.

36. (a) 0,8413; (b) 3; (c) 0,004; (d) 20,260.

38. (a) $1 - \dfrac{1}{X_1}$; (b) $1 - \dfrac{1}{X_1}$.

40. (a) $-$; (b) 0; (c) $\dfrac{2\sigma^4}{n_1 + n_2}$.

42. (a) $\dfrac{\bar{X}}{r}$; (b) $\dfrac{\theta(\theta-1)}{nr}$; (c)

$$\theta_{VM} \pm z_{\alpha/2}\left[\frac{\theta_{VM}(\theta_{VM}-1)}{nr}\right]^{1/2}$$

(d) 1,2 $\pm$ 0,043.

44. $\dfrac{1}{\bar{X}}\{1 \pm 0{,}429\}$ (environ).

46. (a) $\dfrac{n}{\sum_{i=1}^{n}|X_i|}$; (b) $\left[\dfrac{2n}{\sum_{i=1}^{n}X_i^2}\right]^{1/2}$.

48. (a) $\frac{1}{n}\sum_{i=1}^{n}X_i^3$; (b) $\dfrac{\theta^2}{n}$; (c)

$$\left[\frac{2\sum_{i=1}^{n}X_i^3}{\chi^2_{\alpha/2,2n}}, \frac{2\sum_{i=1}^{n}X_i^3}{\chi^2_{1-\alpha/2,2n}}\right]$$

(d) 385.

50. (a) 0,295 $\pm$ 0,0089 (environ); (b) 5.

52. (a) $-1 - \dfrac{n}{\sum_{i=1}^{n}\ln(2-X_i)}$.

Chapitre 6

2. (a) (i) $d^2 = 4 \leq 7{,}815$; on accepte le modèle; (ii) $d^2 \simeq 55{,}63 > 9{,}488$; on rejette le modèle; (iii) $d^2 \simeq 3{,}16 \leq 7{,}815$; on accepte le modèle; (b) U$[0, 16]$, car c'est la seule distribution bornée des deux côtés.

4. (a) $d^2 \simeq 1{,}93 \leq 6{,}251$; on accepte le modèle; (b) $z_0 \simeq -0{,}8165 \geq -1{,}282$; on ne rejette pas H_0.

6. X_A (respectivement X_B) = rendement (en %) avec le catalyseur A (respectivement B). On suppose que $X_A \approx \mathrm{N}(\mu_A, \sigma_A^2)$ et $X_B \approx \mathrm{N}(\mu_B, \sigma_B^2)$ sont des variables aléatoires indépendantes. On teste d'abord l'égalité des variances. Puisque

$$f_0 \simeq 0{,}866 \in [0{,}143; 5{,}99]$$

on ne rejette pas H_0. Ensuite, on teste l'égalité des moyennes. On calcule $s_p^2 \simeq 2{,}363$ et alors $t_0 \simeq -4{,}68$. Étant donné que

$$|t_0| > t_{0{,}025;10} \simeq 2{,}226 > t_{0{,}025;11}$$

on rejette H_0.

8. (a) $d^2 = 5{,}\bar{5} \leq 7{,}815$; on accepte l'hypothèse d'indépendance des variables; (b) [8,91 %; 31,09 %] (environ); (c) $d^2 = 4 \leq 7{,}815$; on accepte H_0.

10. (a) X = diamètre des axes; on suppose que $X \approx \mathrm{N}(\mu, \sigma^2)$, où μ et σ^2 sont inconnus; on veut tester H_0: $\sigma^2 = 1{,}44$ contre H_1: $\sigma^2 > 1{,}44$; (b) on rejette H_0 ssi $S^2 > 2{,}71$ (environ); (c) 0,16; (d) 15.

12. (a) X_1 (respectivement X_2) = quantité de combustible synthétique par kilogramme d'hydrogène utilisé avec l'ancien (respectivement le nouveau) procédé; on suppose que $X_i \approx \mathrm{N}(\mu_i, \sigma_i^2)$, où tous les paramètres sont inconnus, et que X_1 et X_2 sont indépendantes; on veut tester H_0: $\sigma_1^2 = \sigma_2^2$ contre H_1: $\sigma_1^2 > \sigma_2^2$; puisque $f_0 \simeq 21{,}52 \geq 3{,}10$, on rejette H_0; (b) on veut tester H_0: $\mu_1 = \mu_2$ contre H_1: $\mu_1 < \mu_2$; on calcule $t_0^* \simeq -7{,}28 < -t_{0{,}05;9}$ (et $-t_{0{,}05;20}$); alors on rejette H_0; (c) on a maintenant $f_0 \simeq 2{,}13 \leq 3{,}01$; alors on ne peut pas rejeter H_0; (d) l'intervalle de confiance pour $\mu_1 - \mu_2$ est $[-4{,}10; -3{,}30]$ (environ); puisque $-3{,}30 < 0$, on peut conclure que $\mu_1 < \mu_2$.

14. (a) $X \sim \mathrm{Poi}(\lambda)$ implique que $\bar{X} \approx \mathrm{N}\big(\lambda, \frac{\lambda}{n}\big)$ par le théorème central limite; (b) $C = \lambda_0 + z_\alpha \sqrt{\frac{\lambda_0}{n}}$; (c) $\bar{x} = 1{,}2 \leq C \simeq 1{,}233$; on ne rejette pas H_0; (d) 0,5832; (e) 298 (ou 297).

16. (a) $|t_0| \simeq 3{,}46 > 1{,}833$; on rejette H_0 (test avec observations appariées); (b) (i) $f_0 \simeq 0{,}87 \in [0{,}25; 4{,}03]$; on ne rejette pas H_0; (ii) $|t_0| \simeq 2{,}01 \leq 2{,}101$; on ne rejette pas H_0; (c) (i) $z_0 \simeq -0{,}973 \geq -2{,}326$; on ne rejette pas H_0; (ii) 0,8665; (iii) 61; (d) (i) $206 \pm 6{,}66$ (environ); (ii) [47,64; 556,59] (environ).

18. (a) Soit X le nombre d'articles défectueux dans un échantillon; on a: $X_A \sim \mathrm{B}(n_A, p_A)$ ($\approx \mathrm{N}(n_A p_A, n_A p_A (1 - p_A))$) et $X_B \sim \mathrm{B}(n_B, p_B)$; on suppose que X_A

et X_B sont indépendantes; on veut tester H_0: $p_A = p_B$ contre H_1: $p_A > p_B$; (b) $z_0 \simeq 0{,}507 \leq 1{,}282$; on ne rejette pas H_0; (c) 11.

20. (a) $t_0 \simeq 3{,}35 > 2{,}602 > t_{0{,}01;19}$; on peut rejeter H_0: $\mu = 30$ % et entreprendre l'exploitation de la mine; (b) 0,536; (c) (28 ou) 29.

22. (a) $\theta_{VM} = \bar{X}/n = 1/20$; (b) $d^2 \simeq 0{,}112 \leq 3{,}841$; on accepte le modèle.

24. (a) $z_0 \simeq -2{,}26 < -1{,}645$; on accepte H_1: $\mu_A < \mu_B$; (b) 0,2776; (c) 18.

26. (a) (i) $|t_0| \simeq 0{,}57 \leq 2{,}528 < t_{0{,}025;19}$; on ne rejette pas H_0; (ii) $w_0^2 \simeq 21{,}35 \in [8{,}91; 32{,}85]$; on ne rejette pas H_0; (b) $d^2 = 1{,}2 \leq 6{,}251$; on accepte H_0.

28. (a) H_0: $\mu = 330$ contre H_1: $\mu < 330$; $t_0 \simeq -1{,}476 < -1{,}383$; on rejette H_0; (b) $w_0^2 \simeq 22{,}62 > 16{,}92$; on accepte H_1: $\sigma^2 > 30^2$; (c) $z_0 \simeq -2{,}34 < -1{,}282$; on rejette H_0; (d) 0,4 (environ).

30. (a) $t_0 \simeq 1{,}815 \leq 3{,}143$; on ne rejette pas H_0: $\mu = 0{,}07$; (b) en utilisant la formule dans le cas où σ est connu, on trouve que $n_{\min} = 63$ ou 64; (c) la formule avec σ connu donne environ 0,9985.

32. (a) $d^2 = 4{,}724 \leq 14{,}68$; on conclut que le générateur fonctionne bien; (b) 0,4975.

34. (a) $w_0^2 \simeq 23{,}5 \leq 30{,}144$; on ne rejette pas H_0: $\sigma^2 = (1{,}5)^2$; (b) 0,21 (environ); (c) 39.

Chapitre 7

2. (a) $\hat{\beta}_0 \simeq 52{,}38$; $\hat{\beta}_1 \simeq -1{,}383$; $\hat{\sigma}^2 \simeq 1{,}89$; $f_0 \simeq 375{,}93$; (b) $\hat{N}(40) \simeq -2{,}94 < 0 \Rightarrow 0$; (c) $[0; 3{,}43]$ (environ).

4. (a) $\hat{\beta}_0 \simeq 8{,}31$; $\hat{\beta}_1 \simeq -1{,}08$; $\hat{\sigma}^2 \simeq 0{,}166$; (b) $f_0 \simeq 113{,}55 > 5{,}32$; on rejette H_0.

6. (a) $\hat{\alpha} \simeq 14{,}76$; $\hat{\gamma} \simeq 2{,}65$; (b) $f_0 \simeq 13{,}4$; $R^2 \simeq 0{,}81$; (c) $[12{,}3; 44{,}3]$ (environ).

8. (a) $r_{X,Y} \simeq 0{,}83 \Rightarrow$ oui; (b) 230; (c) (i) les observations de U et de V sont les mêmes, seul l'ordre des observations est différent; (ii) $\hat{\beta}_0 \simeq 6{,}4$; $\hat{\beta}_1 \simeq 0{,}81$.

10. (a) 0,94; (b) 82,6.

12. (a) $\bar{Y}$; (b) $\bar{Y} - \hat{\beta}_1 \sum_{i=1}^{n} \dfrac{x_i^2}{n}$.

14. (a) $0{,}8\bar{3}$; (b) $1{,}\bar{6}$.

16. (a) $Y' = \ln Y$; $x' = \ln x$; $\beta_0' = \ln \beta_0$; $\beta_1' = \beta_0 \beta_1$; (b) $12 > F_{0{,}01;1,8} = t_{0{,}005;8}^2 \simeq 11{,}26$; on rejette H_0.

Chapitre 8

2. $r(x) = [x(1 - \ln x)]^{-1}$, pour $1 \leq x \leq e$. On trouve que $r'(x) > 0$ dans l'intervalle $[1, e]$. De là, la distribution est IFR.
4. $R(t) = (e^{-t} - e^{-2t})/(2t)$, pour $t \geq 0$.
6. $\frac{71}{19}e^{-2}$.
8. $r_X(k) = \frac{1}{N-k+1}$. La distribution est IFR.
10. (a) $2\lambda/3$; (b) $\lambda/2$.
12. $1/3$.
14. $R(t) = 1 - (1 - e^{-\lambda_C t})\left\{1 - \left[1 - (1 - e^{-\lambda_A t})^3\right] e^{-\lambda_B t}\right\}$.
16. $\simeq 0{,}3710$.
18. $0{,}996303 \leq R(t_0) \leq 0{,}997107$ (approximativement). La réponse exacte est 0,996327.
20. (a) $p^3(3 - 2p)$; (b) $e^{-3\theta t}(3 - 2e^{-\theta t})$, pour $t \geq 0$.

Chapitre 9

2. $\{N^*(t), t \geq 0\}$ est un processus de Poisson de taux $\lambda = 1$. Donc, c'est effectivement un processus de naissance pur.
4. (a) 1/3; (b) 1/3; (c) 2/3.
6. (a) 2,5; (b) 2,72; (c) 2,89.
8. $1/16$.
10. 0,079.
12. $\left(1 - e^{-2\mu}\right)/(2\mu)$.
14. On définit les états:

0 : le système est vide
n_0 : il y a n clients dans le système, dont un qui attend de savoir si le serveur peut le servir ou non
n : il y a n clients dans le système, dont un qui est en train d'être servi

pour $n = 1, 2, 3$. Les équations d'équilibre sont:

état j	taux de départ de j = taux d'arrivée à j
0	$\lambda\pi_0 = \mu_0 p\pi_{1_0} + \mu\pi_1$
1_0	$(\lambda+\mu_0)\pi_{1_0} = \lambda\pi_0 + \mu_0 p\pi_{2_0} + \mu\pi_2$
2_0	$(\lambda+\mu_0)\pi_{2_0} = \lambda\pi_{1_0} + \mu_0 p\pi_{3_0} + \mu\pi_3$
3_0	$\mu_0\pi_{3_0} = \lambda\pi_{2_0}$
1	$(\lambda+\mu)\pi_1 = \mu_0(1-p)\pi_{1_0}$
2	$(\lambda+\mu)\pi_2 = \mu_0(1-p)\pi_{2_0}$
3	$\mu\pi_3 = \mu_0(1-p)\pi_{3_0}$

16. (a) $\mu/[4(\mu+\lambda)]$; (b) $1-\{\mu/[2(\mu+\lambda)]\}$.
18. 207/256.
20. 0,77.

D

Réponses des questions à choix multiple

Chapitre 1

1 b; 2 e; 3 b; 4 c; 5 a; 6 d; 7 b; 8 c; 9 e; 10 a.

Chapitre 2

1 c,b,b,b,b,e; 2 c; 3 d; 4 d; 5 d; 6 c; 7 d; 8 e; 9 a; 10 c; 11 c; 12 b; 13 c; 14 a.

Chapitre 3

1 d; 2 a; 3 d; 4 b; 5 c; 6 c; 7 a; 8 a; 9 e; 10 a; 11 b; 12 e; 13 a; 14 d; 15 a; 16 c; 17 c; 18 e; 19 c; 20 c.

Chapitre 4

1 e; 2 b; 3 c; 4 a; 5 a; 6 e; 7 c; 8 d; 9 b; 10 c; 11 d; 12 c; 13 c; 14 e; 15 d; 16 c; 17 b; 18 e; 19 d; 20 c.

Chapitre 5

1 a; 2 e; 3 d; 4 d; 5 b; 6 b; 7 d,c,e; 8 e,c,d; 9 e,e,c; 10 d,b,c; 11 c,e; 12 d,a.

Chapitre 6

1 d,b,d; 2 a,b; 3 e,a,a; 4 b,c,d; 5 d,b,d; 6 b; 7 a,b,a,c; 8 d,d; 9 a,c,a.

Chapitre 7

1 e,c,e; 2 e; 3 d,c,d; 4 b,c,e.

Chapitre 8

1 d; 2 e; 3 c; 4 a; 5 b; 6 c; 7 c; 8 c; 9 d; 10 d.

Chapitre 9

1 b; 2 d; 3 e; 4 b; 5 d; 6 a; 7 c; 8 d; 9 b; 10 c.

Bibliographie

1. Barnes, J. Wesley, *Statistical Analysis for Engineers: A Computer-Based Approach*, Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
2. Bowker, Albert H., et Lieberman, Gerald J., *Engineering Statistics*, Prentice-Hall, Englewood Cliffs, New Jersey, 1959.
3. Breiman, Leo, *Probability and Stochastic Processes: With a View Toward Applications*, Houghton Mifflin, Boston, 1969.
4. Cartier, Jacques, Parent, Régis, et Picard, Jean-Marc, *Inférence Statistique*, Éditions Sciences et Culture, Montréal, 1976.
5. Chung, Kai Lai, *Elementary Probability Theory with Stochastic Processes*, Springer-Verlag, New York, 1975.
6. Clément, Bernard, *Analyse Statistique*, Tomes I et II, Notes de cours, École Polytechnique de Montréal, 1988.
7. Dodge, Yadolah, *Analyse de régression appliquée*, Dunod, Paris, 1999.
8. Dougherty, Edward R., *Probability and Statistics for the Engineering, Computing, and Physical Sciences*, Prentice-Hall, Englewood Cliffs, New Jersey, 1990.
9. Feller, William, *An Introduction to Probability Theory and Its Applications*, Volume I, 3^{e} édition, Wiley, New York, 1968.
10. Feller, William, *An Introduction to Probability Theory and Its Applications*, Volume II, 2^{e} édition, Wiley, New York, 1971.
11. Hastings, Kevin J., *Probability and Statistics*, Addison-Wesley, Reading, Massachusetts, 1997.
12. Hines, William W., et Montgomery, Douglas C., *Probability and Statistics in Engineering and Management Science*, 3^{e} édition, Wiley, New York, 1990.
13. Hogg, Robert V., et Craig, Allen T., *Introduction to Mathematical Statistics*, 3^{e} édition, Macmillan, New York, 1970.
14. Hogg, Robert V., et Tanis, Elliot A., *Probability and Statistical Inference*, 6^{e} édition, Prentice Hall, Upper Saddle River, New Jersey, 2001.
15. Krée, Paul, *Introduction aux Mathématiques et à leurs Applications Fondamentales*, Dunod, Paris, 1969.
16. Lapin, Lawrence L., *Probability and Statistics for Modern Engineering*, 2^{e} édition, PWS-KENT, Boston, 1990.

17. Leon-Garcia, Alberto, *Probability and Random Processes for Electrical Engineering*, 2^{e} édition, Addison-Wesley, Reading, Massachusetts, 1994.
18. Lindgren, Bernard W., *Statistical Theory*, 3^{e} édition, Macmillan, New York, 1976.
19. Maksoudian, Y. Leon, *Probability and Statistics with Applications*, International, Scranton, Pennsylvania, 1969.
20. Massey, F. J. Jr., The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American Statistical Association*, vol. 46, p. 68-78, 1951.
21. Miller, Irwin, Freund, John E., et Johnson, Richard A., *Probability and Statistics for Engineers*, 4^{e} édition, Prentice-Hall, Englewood Cliffs, New Jersey, 1990.
22. Montgomery, Douglas C., Peck, Elizabeth A., et Vining, G. Geoffrey, *Introduction to Linear Regression Analysis*, 3^{e} édition, Wiley, New York, 2001.
23. Papoulis, Athanasios, *Probability, Random Variables, and Stochastic Processes*, 3^{e} édition, McGraw-Hill, New York, 1991.
24. Roberts, Richard A., *An Introduction to Applied Probability*, Addison-Wesley, Reading, Massachusetts, 1992.
25. Ross, Sheldon M., *Introduction to Probability and Statistics for Engineers and Scientists*, Wiley, New York, 1987.
26. Ross, Sheldon M., *Introduction to Probability Models*, 7^{e} édition, Academic Press, San Diego, 2000.
27. Shapiro, S. S., et Wilk, M. B., An analysis of variance test for normality (complete samples), *Biometrika*, vol. 52, p. 591-611, 1965.
28. Spiegel, Murray R., *Theory and Problems of Advanced Calculus*, Schaum's Outline Series, McGraw-Hill, New York, 1973.
29. Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L., et Ye, Keying, *Probability and Statistics for Engineers and Scientists*, 7^{e} édition, Prentice Hall, Upper Saddle River, New Jersey, 2002.

Index

Probabilités, statistique et applications

Probabilités, statistique et applications présente toute la théorie essentielle des probabilités et de la statistique; il constitue un excellent choix comme manuel de classe, peu importe la discipline du génie ou des sciences appliquées dans le cadre de laquelle la matière est enseignée. Après une révision des éléments de base du calcul différentiel et intégral, l'ouvrage traite des principaux résultats des probabilités et de la statistique mathématique ainsi que des applications de la théorie. En effet, l'ouvrage privilégie l'application plutôt que les détails mathématiques, car pour tout étudiant, peu importe son niveau et sa formation préalable, la véritable maîtrise des probabilités et de la statistique passe par la résolution d'un grand nombre d'exercices, de types et de degrés de difficulté variés. C'est dans cette optique qu'une part très importante de l'ouvrage est réservée aux exercices; le livre en regroupe près de 600, dont 320 sont résolus.

S'il s'adresse d'abord aux étudiants de premier cycle en génie et en sciences appliquées, *Probabilités, statistique et applications* peut aussi servir de manuel de référence pour les étudiants aux études supérieures dont les projets de recherche exigent des connaissances de base en probabilités et en statistique. En outre, les deux derniers chapitres de l'ouvrage, qui traitent de fiabilité et de modèles de files d'attente, en font un outil particulièrement intéressant pour les étudiants en informatique et en génie industriel.

Mario Lefebvre est professeur titulaire au Département de mathématiques et de génie industriel de l'École Polytechnique de Montréal, où il a coordonné l'enseignement de cours de probabilités et statistique au premier cycle universitaire pendant plusieurs années, tout en enseignant les processus stochastiques aux cycles supérieurs. Il détient un baccalauréat et une maîtrise en mathématiques de l'Université de Montréal ainsi qu'un doctorat en mathématiques de l'Université de Cambridge, en Angleterre.

ISBN : 978-2-553-01554-0

9 782553 015540

PRESSES INTERNATIONALES POLYTECHNIQUE

www.polymtl.ca/pub