Collection
Statistique
et probabilités
appliquées

Michel Lejeune

Statistique

La théorie et ses applications

Deuxième édition

Plus de 150 exercices corrigés

Springer

Statistique
La théorie et ses applications

Deuxième édition
avec exercices corrigés

Springer
Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Michel Lejeune

Statistique

La théorie et ses applications

Deuxième édition
avec exercices corrigés

Springer

Michel Lejeune
Professeur émérite
Université de Grenoble 2
IUT 2 département statistique
BP 47
38040 Grenoble cedex 9

ISBN : 978-2-8178-0156-8 Springer Paris Berlin Heidelberg New York

Springer-Verlag France est membre du groupe Springer Science + Business Media

Maquette de couverture : Jean-François Montmarché
© détail du tableau " Chrysler impressions " de Philippe Lejeune (1990)

Collection
Statistique et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

Comité éditorial :

Dans la même collection :

– *Statistique. La théorie et ses applications*
Michel Lejeune, avril 2004

– *Optimisation appliquée*
Yadolah Dodge, octobre 2004

– *Le choix bayésien. Principes et pratique*
Christian P. Robert, novembre 2005

– *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique*
Eva Cantoni, Philippe Huber, Elvezio Ronchetti, novembre 2006

– *Régression. Théorie et applications*
Pierre-André Cornillon, Éric Matzner-Løber, janvier 2007

– *Le raisonnement bayésien. Modélisation et inférence*
Éric Parent, Jacques Bernier, juillet 2007

– *Premiers pas en simulation*
Yadolah Dodge, Giuseppe Melfi, juin 2008

– *Génétique statistique*
Stephan Morgenthaler, juillet 2008

– *Pratique du calcul bayésien*
Jean-Jacques Boreux, Éric Parent, 2009

– *Maîtriser l'aléatoire*
Eva Cantoni, Philippe Huber, Elvezio Ronchetti, septembre 2009

À paraître :

– *Le logiciel R*
Pierre Lafaye de Micheaux, Rémi Drouilhet, Benoit Liquet, 2010

AVANT-PROPOS

L'objectif de cet ouvrage est de rendre accessibles les fondements théoriques de la statistique à un public de niveau mathématique moyen : étudiants du premier cycle des filières scientifiques, élèves ingénieurs, chercheurs dans les domaines appliqués (économie, gestion, biologie, médecine, géographie, sciences de la vie, psychologie...) et, plus généralement, tous les chercheurs désireux d'approfondir leur compréhension des résultats utilisés dans la pratique. Pour ces derniers un minimum de connaissance de l'arrière-plan théorique apportera une vision plus claire et plus critique des méthodes qu'ils emploient et permettra d'éviter bien des écueils.

Les prérequis principaux sont la maîtrise de la dérivation, de l'intégration et de bases minimales du calcul des probabilités. Sur le plan purement mathématique, nous pensons que l'essentiel de l'exposé est accessible à quiconque aurait parfaitement assimilé le programme d'un bac scientifique. Il reste cependant quelques notions qui ne sont abordées qu'en premier cycle supérieur, notamment les approximations par développement de Taylor, les développements en série entière, les fonctions de plusieurs variables (dérivation et intégration) et, très marginalement, le calcul matriciel. Mais ces notions n'interviennent le plus souvent que dans les aspects techniques de démonstration, ce qui ne devrait pas nuire à la compréhension des concepts. Pour satisfaire la curiosité de mathématiciens qui voudraient, par la lecture de cet ouvrage, s'initier sans peine à la science statistique, mention sera faite ici ou là de résultats ou démonstrations exigeant des connaissances plus approfondies d'analyse. Ces éléments seront consignés en petits caractères, généralement dans des «notes» détachées que l'on pourra ignorer totalement. Quelques exercices plus difficiles, repérés par un astérisque, leur sont également proposés.

Notons que les premiers chapitres concernent la théorie des probabilités qui, toutefois, est abordée non comme une fin en soi mais de façon simplifiée dans la perspective de ce qui est nécessaire pour la théorie statistique de l'estimation et des tests.

Pour atteindre l'objectif fixé nous avons pris le parti de toujours privilégier la facilité de compréhension au détriment éventuel de la pureté formelle (si tant est qu'elle existe). Nous sommes d'avis que trop de formalisme nuit à l'assimilation des concepts et qu'il faut s'efforcer sans cesse de s'en tenir à un niveau compatible avec celui des connaissances du public visé. Ceci a été un souci constant dans la rédaction. Cela ne signifie pas que nous ayons renoncé à la rigueur du propos, c'est-à-dire à la cohérence des éléments apportés tout au long de l'ouvrage.

Par ailleurs, nous faisons partie de ceux qui pensent que la statistique ne relève pas uniquement de la mathématique qui n'est qu'un instrument. Sa

raison d'être consiste à appréhender le monde réel à partir des observations que l'on en fait. C'est pourquoi la discipline est rangée dans le domaine des mathématiques appliquées, ce terme ne devant pas, à notre sens, rester un vain mot. Fidèle à cette vision nous avons tenté de commenter le plus largement possible les concepts et résultats de façon concrète pour montrer leur utilité dans l'approche du réel. Dans les chapitres débouchant immédiatement sur des méthodes usuelles nous avons également introduit des exercices «appliqués» pour illustrer l'intérêt et la mise en oeuvre des principes théoriques. L'ouvrage n'est donc pas uniquement un traité mathématique. Cela a motivé le choix de son sous-titre « La théorie et ses applications» pour marquer la distinction, même si son objectif premier reste l'exposé de la théorie.

L'essentiel de l'apport de cette nouvelle édition est constitué des corrigés détaillés des exercices proposés. Cette demande m'a été faite de façon récurrente et il est vrai que ces corrigés doivent permettre d'améliorer nettement l'assimilation de la matière.

Je remercie mes collègues Alain Latour et Pierre Lafaye de Micheaux pour leur aide technique précieuse ainsi qu'Alain Catalano, Yves-Alain Gerber, Jérôme Hennet, Alexandre Junod, Julien Junod, Vincent Voirol et Mathieu Vuilleumier pour leurs appréciations.

J'adresse des remerciements particuliers à Yadolah Dodge, directeur de cette collection « Statistique et probabilités appliquées », sans les encouragements duquel cet ouvrage n'aurait sans doute pas abouti.

Michel Lejeune *Grenoble, juin 2010*

Table des matières

Chapitre 1

Variables aléatoires

1.1 Notion de variable aléatoire

La théorie des probabilités a pour objet l'étude des phénomènes aléatoires ou du moins considérés comme tels par l'observateur. Pour cela on introduit le concept d'*expérience*[1] *aléatoire* dont l'ensemble des résultats possibles constitue l'*ensemble fondamental,* noté habituellement Ω. On parle de *variable aléatoire* (abréviation : v.a.) lorsque les résultats sont numériques, c'est-à-dire que Ω est identique à tout ou partie de l'ensemble des nombres réels $\mathbb{R}$.

On distingue habituellement :

- les *variables aléatoires discrètes* pour lesquelles l'ensemble Ω des résultats possibles est un ensemble discret de valeurs numériques $x_1, x_2, \ldots, x_n, \ldots$ fini ou infini (typiquement : l'ensemble des entiers naturels) ;
- les *variables aléatoires continues* pour lesquelles l'ensemble Ω est tout $\mathbb{R}$ (ou un intervalle de $\mathbb{R}$ ou, plus rarement, une union d'intervalles).

On peut concevoir des variables mixtes, mais nous ne traiterons pas, sauf exception, ces cas particuliers.

Dans toute expérience aléatoire on est amené à s'intéresser à des ensembles de résultats, donc des parties de Ω, que l'on appelle *événements*, les résultats formant eux-mêmes des *événements élémentaires*. Dans le cas d'une v.a. les événements sont des parties de $\mathbb{R}$, le plus souvent des intervalles. Par exemple on s'intéressera au fait qu'un assuré occasionne un sinistre de coût supérieur à 1000 euros au cours d'une année.

Dès lors il reste à construire un modèle probabiliste pour l'ensemble fondamental considéré. Ceci ne pose pas de problème pour le cas d'une v.a. discrète. En effet il suffit de définir les probabilités de chaque résultat $x_1, x_2, \ldots, x_n, \ldots,$

[1] Le terme est trop restrictif pour prendre en compte la variété des phénomènes étudiés. On trouvera une brève discussion à ce propos au début du chapitre 5.

à partir de quoi on peut, par les règles élémentaires des probabilités, calculer la probabilité de tout événement (en sommant celles des résultats appartenant à l'événement). De plus toute partie de Ω est un événement. C'est la présentation que l'on trouve généralement dans les traités élémentaires.

Pour une v.a.continue les choses sont plus délicates. En effet un point de $\mathbb{R}$ est un intervalle de longueur nulle et la probabilité associée à tout point est elle-même nulle. On ne peut donc «probabiliser » $\mathbb{R}$ à partir de probabilités associées à chacun de ses éléments. En fait les probabilités doivent être attribuées aux événements. De plus, contrairement au cas discret, l'ensemble des parties de $\mathbb{R}$ est trop vaste pour constituer un ensemble d'événements tous probabilisables et l'on doit se restreindre à certaines parties (voir plus loin la note 1.2). Ceci n'a toutefois aucune incidence sur le plan pratique tant il est vrai que les parties de $\mathbb{R}$ qui sont exclues ne sont que des curiosités mathématiques. Par souci d'homogénéité, dans le cas discret on considère la probabilisation de Ω à partir de l'ensemble $\mathcal{E}$ des événements, comme dans le cas continu.

Soit, donc, l'ensemble $\mathcal{E}$ des événements construit à partir de Ω, on appelle *mesure de probabilité* une fonction P qui à tout événement E fait correspondre un nombre $P(E)$ entre 0 et 1 que l'on appellera la probabilité de l'événement E (cette fonction doit en outre vérifier certains axiomes, voir ci-après). Pour une variable aléatoire on parlera plutôt de *la loi de la variable aléatoire* ou encore, de sa *distribution*, par emprunt à la statistique descriptive.

Par commodité on désigne une variable aléatoire par une **lettre majuscule** symbolique et on écrit simplement un événement sous la forme usuelle des notations mathématiques. Ainsi, si X désigne la variable aléatoire «durée de vie en années d'un aspirateur donné», $(X < 3)$ dénotera l'événement «l'aspirateur a une durée de vie inférieure à 3 ans». La probabilité associée à cet événement pourra s'écrire $P(X < 3)$. Cette commodité pourra parfois prêter à confusion et il sera toujours utile de garder à l'esprit son caractère conventionnel. Ainsi dans notre exemple $P(X < 3)$ n'est rien d'autre que la mesure de probabilité associée à l'intervalle $]-\infty, 3[$, soit $P(\,]-\infty, 3[)$. Dans sa forme la plus générale un événement pourra s'écrire $(X \in A)$ où A est une partie de $\mathbb{R}$.

Rappelons succinctement les principales propriétés d'une mesure de probabilité.

1. $P(E) \in [0, 1]$ pour tout événement E et $P(\not\leqslant) = 1$
2. $P(\overline{E}) = 1 - P(E)$, $\overline{E}$ étant le complémentaire de E
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2)$, pour tous événements E_1 et E_2 *incompatibles* (i.e. parties disjointes de $\Omega : E_1 \cap E_2 = \emptyset$)
4. $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$ dans le cas général
5. $E_1 \subseteq E_2$ (E_1 inclus dans E_2) $\Longrightarrow P(E_1) \leq P(E_2)$

6. La *probabilité conditionnelle* de E_1 sachant E_2 (pour autant que l'on ait $P(E_2) \neq 0$) est :

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

7. Les événements E_1 et E_2 de probabilités non nulles sont *indépendants* si et seulement si :

$$P(E_1 \cap E_2) = P(E_1)P(E_2)$$

ou, de façon équivalente quand $P(E_2) \neq 0$, $P(E_1|E_2) = P(E_1)$.

La propriété 1 et la propriété 3 généralisée à une suite $E_1, E_2, \ldots, E_n, \ldots$ d'événements deux à deux incompatibles constituent les **axiomes de la théorie des probabilités**. La propriété 7 s'étend à une suite d'événements de la façon suivante : on dit que les événements $E_1, E_2, \ldots, E_n, \ldots$ sont *(mutuellement) indépendants* si, pour tout sous-ensemble de ces événements, la probabilité de leur intersection est égale au produit de leurs probabilités (donc la relation doit être vérifiée pour les événements pris deux à deux, trois à trois, etc.).

Note 1.1 Plus formellement et pour être plus général, on définit une v.a. en partant d'une expérience aléatoire dont l'ensemble fondamental peut être de **nature quelconque.** C'est pour cet ensemble fondamental qu'est définie la mesure de probabilité pour former un *espace probabilisé*. Une v.a. devient alors une **fonction** de Ω dans $\mathbb{R}$, qui affecte donc à chaque résultat possible une valeur numérique. Par exemple, si l'expérience aléatoire consiste à tirer au hasard un individu dans une population, l'ensemble des résultats possibles Ω est l'ensemble des individus de la population. A partir de là on peut observer l'âge de l'individu. Définissant ainsi la v.a. X «âge d'un individu tiré au hasard dans la population» on obtient la probabilité, disons, de l'événement $(18 \leq X \leq 20)$ en calculant sur Ω la probabilité de tirer un individu dont l'âge est compris dans cet intervalle. Plus généralement à tout événement $E \subseteq \mathbb{R}$ sur X on attribue la probabilité de l'événement $X^{-1}(E) = \{\omega \in \Omega \mid X(\omega) \in E\}$ de l'espace probabilisé initial (cet événement correspond à l'ensemble des résultats possibles ω dans Ω qui conduisent par la fonction X à une valeur appartenant à E). Pour des fonctions X extrêmement singulières il se pourrait que $X^{-1}(E)$ ne soit pas un événement probabilisable. On ne considèrera donc que des *fonctions mesurables,* c'est-à-dire telles qu'une probabilité puisse être affectée à tout E. En pratique toutes les fonctions utilisées sont mesurables et nous ignorerons ce problème dans cet ouvrage.

Ce formalisme n'a d'intérêt que s'il pré-existe, en amont du phénomène numérique observé, des événements de probabilités connues ou facilement calculables. C'est ainsi dans notre exemple : tous les individus ont la même probabilité d'être tirés, égale à $1/N$ où N est le nombre total d'individus. Alors la

probabilité d'un événement E pour X est $1/N$ fois le nombre d'individus dont l'âge est dans E, donc la proportion d'individus dont l'âge est dans E .

Illustrons encore cela par un autre exemple fondé sur le jeu de cartes du bridge. Un joueur donné reçoit 13 cartes parmi les 52 cartes au total. Les cartes étant distribuées au hasard toutes les $\binom{52}{13}$ combinaisons de 13 cartes sont *a priori* équiprobables. Pour évaluer son «jeu» un joueur utilise le système de points classique suivant : un as vaut 4 points, un roi 3, une dame 2 et un valet 1. Ainsi on définit une v.a. X «nombre de points dans son jeu». Pour calculer $P(X = 1)$, par exemple, il suffit (moyennant une bonne maîtrise de l'analyse combinatoire !) de dénombrer les combinaisons de 13 cartes ayant un seul valet et ni as ni roi ni dame (il y en a un nombre $\binom{4}{1}\binom{36}{12}$). La probabilité est alors égale au nombre de ces jeux divisé par le nombre total de combinaisons. On voit comment, dans un tel cas, pour trouver la loi de X il est nécessaire de remonter à l'expérience initiale du tirage au hasard de 13 cartes à laquelle s'applique de façon réaliste le modèle d'équiprobabilité.

1.2 Fonction de répartition

La fonction de répartition est l'instrument de référence pour définir de façon unifiée la loi de probabilité d'une variable aléatoire qu'elle soit discrète ou continue. Si cette fonction est connue, il est possible de calculer la probabilité de tout intervalle et donc, en pratique, de tout événement. C'est pourquoi c'est elle qui est donnée dans les tables des lois de probabilité.

Définition 1.1 *Soit X une variable aléatoire, on appelle* ***fonction de répartition*** *de X, que l'on note F_X, la fonction définie sur $\mathbb{R}$ par :*

$$F_X(x) = P(X \leq x).$$

La valeur prise par la fonction de répartition au point x est donc la probabilité de l'événement $]-\infty, x]$. En anglais on l'appelle «cumulative distribution function» par analogie avec la notion de fréquence cumulée en statistique descriptive.

Note 1.2 La fonction de répartition est définie pour tout $x \in \mathbb{R}$. La question se pose de savoir si la connaissance de F_X, donc des probabilités de tous les événements de la forme $]-\infty, x]$, suffit pour déterminer la probabilité d'un événement quelconque.

Pour une v.a. discrète, il est clair que par soustraction on peut déterminer la probabilité de chaque valeur possible et, à partir de là, de toutes les parties de Ω par simple sommation. Toutefois dans les traités élémentaires où ne sont abordées que les v.a. discrètes, l'utilisation de la fonction de répartition n'est pas nécessaire, puisque l'on peut se contenter des probabilités individuelles.

Pour les v.a. continues, comme il a été brièvement indiqué plus haut, on ne peut définir une mesure de probabilité sur toutes les parties de $\mathbb{R}$ qui satisfasse aux axiomes de la théorie. On est conduit à se restreindre aux événements appartenant à *la tribu borélienne* de $\mathbb{R}$. Cette tribu est l'ensemble des parties de $\mathbb{R}$ engendrées par les unions, intersections et compléments d'événements (éventuellement en suite infinie) de la forme $(X \leq x)$. On comprend ainsi que F_X permette, en principe, de calculer la probabilité de tout événement. La restriction à la tribu borélienne de $\mathbb{R}$ n'est pas contraignante car elle contient en fait toutes parties concevables de $\mathbb{R}$ (points isolés, intervalles ouverts ou fermés, unions de tels intervalles, etc.). A vrai dire il faut faire preuve de beaucoup d'ingéniosité pour mettre en évidence une partie de $\mathbb{R}$ n'appartenant pas à la tribu borélienne et nous n'aurons pas à nous préoccuper en pratique de cette restriction (tout comme il a été dit dans la note 1.1 qu'on ne se préoccuperait pas de vérifier si une fonction est mesurable).

Propriétés

1. F_X est **non décroissante** puisque, pour $h > 0$, $(X \leq x) \subset (X \leq x+h)$ et donc $P(X \leq x) \leq P(X \leq x + h)$.
2. $F_X(x)$ varie de 0 à 1 quand x varie de $-\infty$ à $+\infty$, sachant que $F_X(x)$ **est une probabilité** cumulée à partir de $-\infty$. On écrira, en bref, $F_X(-\infty) = 0$ et $F_X(+\infty) = 1$.
3. F_X est **continue à droite** en tout x et $F_X(x) - F_X(x^-) = P(X = x)$, où $F_X(x^-)$ dénote la limite à gauche au point x.

Montrons succinctement cette dernière propriété qui, comme nous allons le voir, résulte du fait que l'événement $(X \leq x)$ intervenant dans la définition de $F_X(x)$ inclut la valeur x elle-même (pour des éléments de démonstration plus rigoureux des propriétés énoncées ici, voir les exercices proposés en fin de chapitre). Par définition, on a :

$$F_X(x^-) = \lim_{\varepsilon \to 0, \varepsilon > 0} F_X(x - \varepsilon) = \lim_{\varepsilon \to 0, \varepsilon > 0} P(X \leq x - \varepsilon)\,.$$

Comme tout événement $(X \leq x - \varepsilon)$ ne contient pas x on admettra qu'au passage à la limite on obtient $F_X(x^-) = P(X < x)$. On a également :

$$F_X(x^+) = \lim_{\varepsilon \to 0, \varepsilon > 0} F_X(x + \varepsilon) = \lim_{\varepsilon \to 0, \varepsilon > 0} P(X \leq x + \varepsilon)\,,$$

mais ici $(X \leq x + \varepsilon)$ contient toujours x et, donc, au passage à la limite on obtient $F_X(x^+) = P(X \leq x) = F_X(x)$. Comme les événements $(X < x)$ et $(X = x)$ sont incompatibles, on en déduit :

$$F_X(x) = F_X(x^-) + P(X = x).$$

En résumé, si la valeur x considérée reçoit une probabilité non nulle (cas discret), alors il y a un saut de discontinuité à gauche d'amplitude égale à cette probabilité, sinon F_X est également continue à gauche et donc continue en x. Nous revenons sur ces notions dans les cas particuliers des variables aléatoires discrètes et des variables aléatoires continues.

1.3 Cas des variables aléatoires discrètes

Pour une variable aléatoire discrète X, l'ensemble des valeurs possibles est un ensemble discret, fini ou infini, de points que nous noterons en ordre croissant : $x_1 < x_2 < \cdots < x_i < \cdots$, sans préciser si l'on est dans le cas fini ou dans le cas infini.

En vertu de ce qui vient d'être vu, la fonction de répartition reste constante entre deux valeurs possibles et présente un saut de discontinuité dès qu'on arrive sur une valeur x_i. En x_i le saut est égal à la probabilité associée à ce point. Immédiatement à gauche de x_i la fonction est égale à $F_X(x_{i-1})$, en x_i et à droite elle est égale à $F_X(x_i)$ (continuité à droite). Cette fonction en escalier s'avère peu maniable et il est plus simple, pour définir la loi de X, de recourir à sa *fonction de probabilité* p_X (appelée aussi fonction de masse de probabilité) qui pour tout x_i $(i = 1, 2, \ldots)$ donne directement sa probabilité $p_X(x_i)$.

Prenons l'exemple du nombre d'appels X arrivant à un standard téléphonique au cours d'une minute, pour lequel un modèle de loi de Poisson de moyenne 10 serait approprié (voir cette loi en section 4.1.7). La variable aléatoire X est définie par :

valeurs possibles	0	1	2	$\cdots$	k	$\cdots$
probabilités associées	e^{-10}	$10e^{-10}$	$\dfrac{10^2e^{-10}}{2}$	$\cdots$	$\dfrac{10^ke^{-10}}{k!}$	$\cdots$

ce qui donne le diagramme en bâtonnets et la fonction de répartition de la figure 1.1.

Le passage de la fonction de répartition à la fonction de probabilité et inversement est :

$$p_X(x_i) = F_X(x_i) - F_X(x_i^-) = F_X(x_i) - F_X(x_{i-1})$$
$$F_X(x) = \sum_{i \mid x_i \leq x} p_X(x_i)$$

1.4 Cas des variables aléatoires continues

Formellement on dira qu'une variable aléatoire X est *continue* s'il existe une fonction f_X non négative telle que, pour tout $x \in \mathbb{R}$, la fonction de répartition

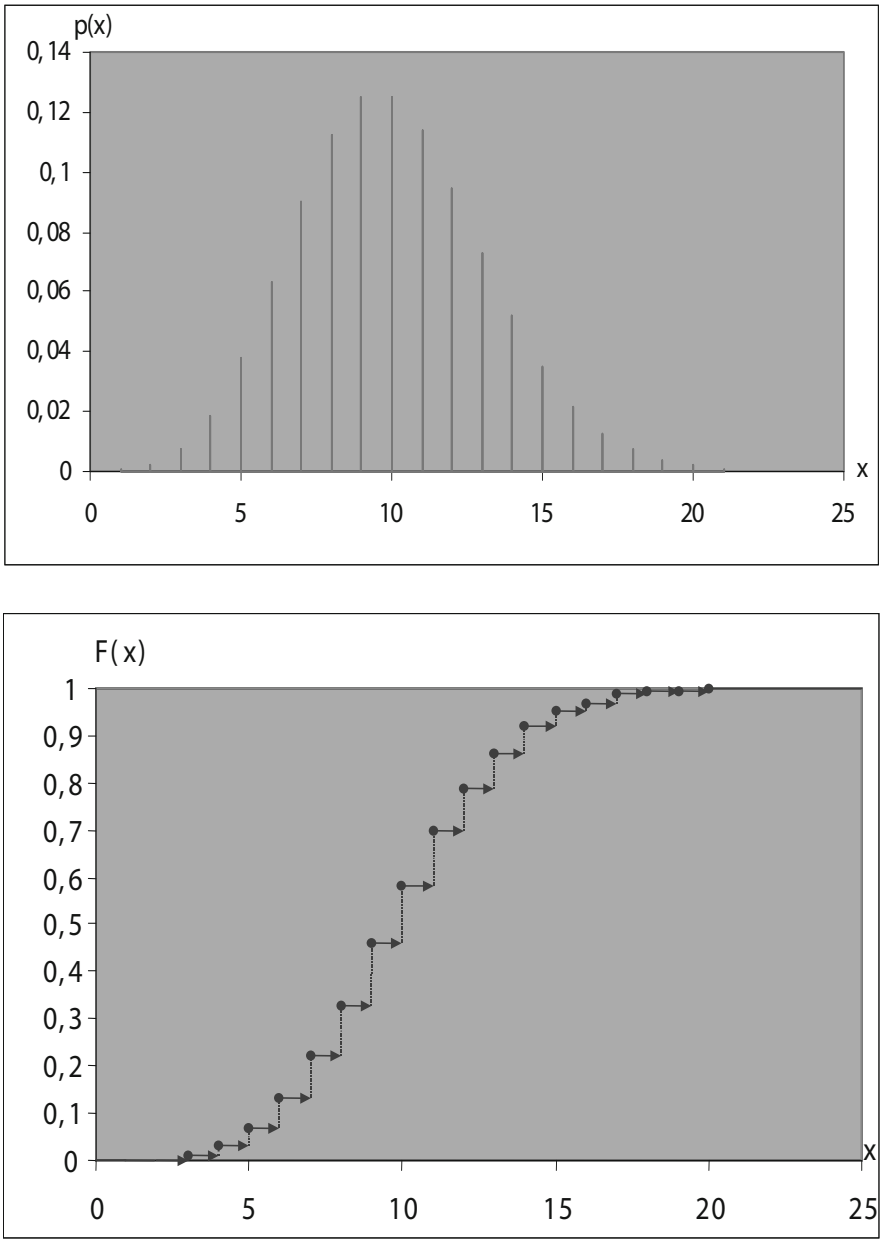


Figure 1.1 - Fonction de probabilité et fonction de répartition de la loi de Poisson de moyenne 10.

puisse s'écrire

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \quad .$$

La fonction f_X est alors appelée *fonction de densité de probabilité* de X ou simplement densité de X. Le fait que F_X s'exprime comme une intégrale implique qu'elle est continue partout et, par conséquent, pour tout x on a $P(X = x) = F_X(x) - F_X(x^-) = 0$. Plus concrètement, chaque point de la droite réelle est immatériel en tant qu'intervalle de longueur nulle et a une probabilité nulle en tant qu'événement, mais peut être caractérisé par une densité de probabilité en ce point. Les événements d'intérêt seront généralement des **intervalles** et pour ceux-ci **il sera indifférent d'y inclure ou non les bornes**.

La fonction de répartition devient alors particulièrement appropriée pour calculer la probabilité de tout intervalle $[a, b]$. En effet, comme on a :

$$(X \leq b) = (X \leq a) \cup (a < X \leq b) \ ,$$

les deux événements à droite étant incompatibles et les signes $<$ pouvant être remplacés par $\leq$, il s'ensuit que :

$$\begin{aligned} P(X \leq b) &= P(X \leq a) + P(a \leq X \leq b) \\ P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \end{aligned}$$

d'où, par la définition même de F_X, les formules fondamentales :

$$\begin{aligned} P(a \leq X \leq b) &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(u)du \, . \end{aligned}$$

On remarquera au passage que, dans le cas discret, les formules se compliquent car selon qu'on inclut ou non les bornes de l'intervalle il faut introduire $F_X(b^-)$ et $F_X(a^-)$.

On admettra que, plus généralement, pour tout événement $(X \in A)$ on a :

$$P(X \in A) = \int_A f_X(x)dx.$$

En général F_X sera dérivable partout sauf peut-être en quelques points qui seront des points de discontinuité pour f_X (d'un point de vue purement mathématique F_X existerait même si f_X était discontinue sur un ensemble dénombrable de points). En un point x où elle est dérivable prenons un intervalle de longueur h centré sur x. La probabilité associée à cet intervalle est alors $P(x - \frac{h}{2} < X < x + \frac{h}{2}) = F_X(x + \frac{h}{2}) - F_X(x - \frac{h}{2})$, d'où :

$$\lim_{h\to 0} \frac{P(x-\frac{h}{2} < X < x+\frac{h}{2})}{h} = F_X^{'}(x) = f_X(x)\,,$$

ce qui justifie l'appellation de densité de probabilité.

Bien que d'un point de vue pratique, pour les modèles de lois continues, ce soit F_X qui soit utile - et c'est bien elle qui est donnée dans les tables - la représentation graphique de f_X est plus parlante car elle met en évidence les zones à plus forte probabilité. Chacun sait interpréter intuitivement, par exemple, la fameuse courbe en cloche du modèle de la loi de Gauss.

A titre illustratif, considérons le jeu de loterie où l'on fait tourner une flèche sur un cadran avec une zone gagnante, et soit la variable X correspondant à l'angle de la flèche, par rapport à une origine déterminée, après expérience. S'il n'y a pas de direction privilégiée la densité de probabilité est la même partout, c'est-à-dire sur l'intervalle $[0\,,\,360]$, et X suit une loi continue uniforme (voir section 4.2.1) sur celui-ci. Les graphes de la densité et de la fonction de répartition sont donnés en figure 1.2. La probabilité d'un intervalle $[a, b]$, correspondant à la surface sous la densité, y est mise en évidence. Notons que F_X est dérivable partout sauf aux bornes du *support* de f_X (on appelle support de f_X l'ensemble des valeurs où elle n'est pas nulle).

Outre qu'elle est une fonction non négative, la densité a les propriétés suivantes :

$$\int_{-\infty}^{+\infty} f_X(x)dx = 1\,,$$
$$\lim_{x\to\pm\infty} f_X(x) = 0\,,$$

la première inégalité découlant du fait que l'intégrale vaut $F_X(+\infty) - F_X(-\infty)$ (voir la propriété n° 2 en section 1.2), la deuxième étant nécessaire (mais non suffisante) pour que l'intégrale converge aux deux bornes.

Note 1.3 Lorsqu'on aborde la théorie des probabilités par la théorie de la mesure, il n'y a pas lieu de faire de distinction entre variables discrètes et variables continues, et donc entre p_X et f_X. Dans les deux cas il s'agit d'une densité par rapport à la mesure générée par F_X.

1.5 Notion essentielle de quantile

Définition 1.2 *On appelle* ***quantile d'ordre*** q *de la variable* X*, où* $q \in [0, 1]$*, la valeur* x_q *telle que* $P(X \leq x_q) = q$ *ou, de même,* $F_X(x_q) = q$*.*

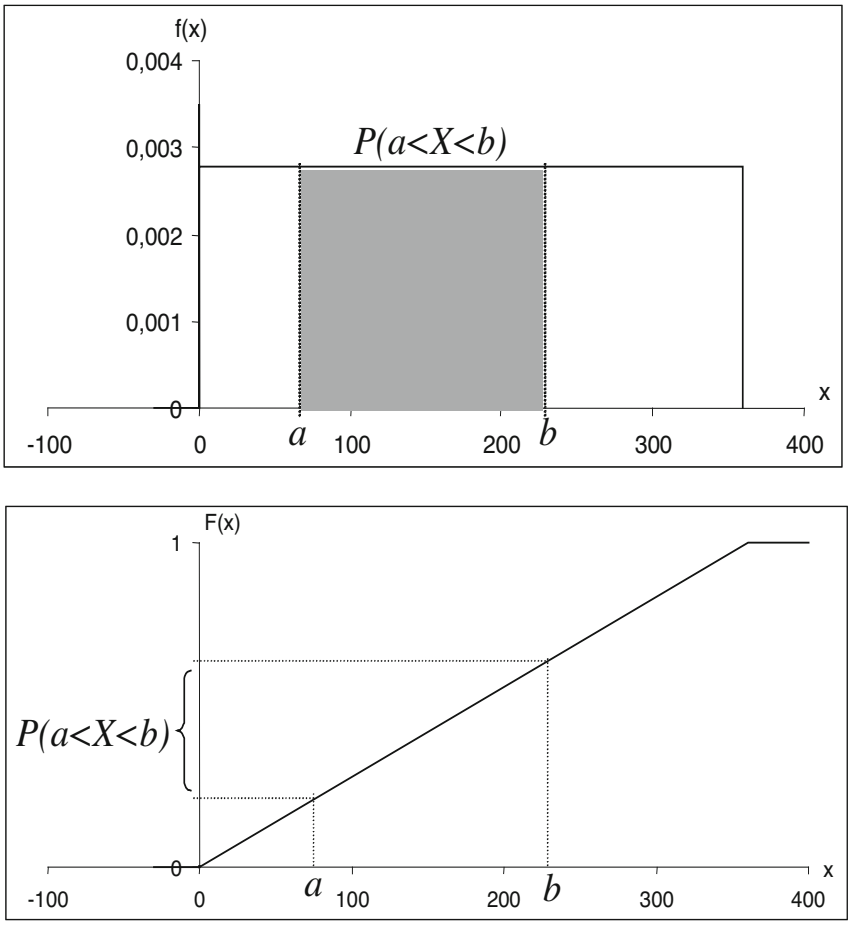


Figure 1.2 - Fonction de densité et fonction de répartition de la loi continue uniforme sur $[0, 1]$.

La notion de quantile (appelée aussi *fractile,* ou *percentile* si exprimée en pour cent) est directement liée à celle de fonction de répartition. Toute valeur de $\mathbb{R}$ peut être vue comme un quantile d'un certain ordre. Cette notion de quantile est essentielle du fait que l'ordre d'un quantile permet de positionner la valeur correspondante sur la distribution considérée. Ainsi le quantile d'ordre 0,5 , alias la *médiane*, est une sorte de centre ou milieu de la distribution. En statistique apparaîtra la nécessité de fixer des limites de plausibilité pour les valeurs d'une loi donnée et l'usage est de prendre pour cela les quantiles $x_{0,025}$ et $x_{0,975}$, soit des valeurs à l'intérieur desquelles la v.a. a une probabilité 0,95 de se trouver.

Dans le cas continu, à tout ordre $q \in [0,1]$ correspond une valeur x_q du fait de la continuité de F_X. Généralement F_X est strictement croissante sur l'ensemble des valeurs de x où $0 < F_X(x) < 1$ et x_q est donc unique pour $q \in]0,1[$.

Dans le cas discret, nous avons vu que F_X est une fonction en escalier et il peut donc y avoir tout un intervalle de valeurs possibles si q correspond au niveau d'une marche de F_X, ou aucune valeur si q est entre deux marches. En pratique on convient de prendre la valeur la plus faible dans le premier cas et d'interpoler linéairement entre les deux valeurs possibles x_i et x_{i+1} telles que $F_X(x_i) < q$ et $F_X(x_{i+1}) > q$ dans le deuxième cas.

1.6 Fonction d'une variable aléatoire

Le problème du passage de la loi d'une v.a. X à la loi d'une fonction $Z = g(X)$ de celle-ci est fréquent. Considérons, par exemple, la v.a. X exprimant la consommation d'une automobile en litres aux 100 kilomètres. A une consommation de x litres/100 km correspond aux Etats-Unis une consommation $z = 235/x$ «miles per gallon» (nombre de miles parcourus avec un gallon d'essence). Ainsi la v.a. X devient une v.a. $Z = 235/X$.

Dans le cas continu la détermination de la loi de la nouvelle v.a. Z passe par sa fonction de répartition $F_Z(z) = P(Z \leq z)$ naturellement définie pour tout $z \in \mathbb{R}$ par la probabilité, pour X, associée à l'ensemble des valeurs x telles que $g(x) \in]-\infty, z]$. En utilisant la symbolique des événements il suffit de résoudre l'événement $(Z \leq z)$ en terme d'événement pour X.

Note 1.4 Rigoureusement, pour que les probabilités des événements sur Z soient calculables il faut que la fonction g soit mesurable (voir note 1.1), c'est-à-dire que pour tout événement E pour Z (donc borélien de $\mathbb{R}$, voir note 1.2) $g^{-1}(E)$ soit un événement pour X (donc un borélien de $\mathbb{R}$). Les fonctions non mesurables ne se rencontrent pas en pratique.

Exemple 1.1 Montrons une fonction strictement croissante, une fonction non monotone et une fonction strictement décroissante.

a) Soit $Z = 2X + 3$:

$$F_Z(z) = P(Z \leq z) = P(2X + 3 \leq z) = P\left(X \leq \frac{z-3}{2}\right) = F_X\left(\frac{z-3}{2}\right).$$

b) Soit $T = X^2$:

$$\begin{aligned} F_T(t) &= P(X^2 \leq t) \\ &= \begin{cases} P(-\sqrt{t} \leq X \leq \sqrt{t}) = F_X(\sqrt{t}) - F_X(-\sqrt{t}) & \text{si } t > 0 \\ 0 \quad \text{si } t \leq 0 \end{cases} . \end{aligned}$$

c) Soit $U = c/X$ où $c > 0$ et X est à valeurs dans $]0, +\infty[$. Pour $u > 0$ on a :

$$F_U(u) = P(U \leq u) = P\left(\frac{c}{X} \leq u\right) = P\left(X \geq \frac{c}{u}\right) = 1 - F_X\left(\frac{c}{u}\right)$$

et $F_U(u) = 0$ pour $u \leq 0$. ■

Si g est strictement croissante comme dans le cas a) ci-dessus, le passage de F_X à F_Z est simple puisque $F_Z(z) = F_X(g^{-1}(z))$. Si g est strictement décroissante comme dans le cas c) on a $F_Z(z) = 1 - F_X(g^{-1}(z))$.

La densité de Z s'obtient simplement par dérivation de F_Z.

Pour les v.a. discrètes, la fonction de répartition, nous l'avons vu, est peu commode et l'on passera par la définition de la fonction de probabilité. Dans les notations du type de celles introduites en début de section 3, l'ensemble des valeurs possibles z_k pour $k = 1, 2, \ldots$ est l'ensemble des valeurs engendrées par $g(x_i)$ pour $i = 1, 2, \ldots$. La probabilité $p_Z(z_k)$ est obtenue en sommant les probabilités $p_X(x_i)$ des valeurs x_i telles que $g(x_i) = z_k$.

1.7 Exercices

Les exercices 1.1 à 1.4 sont d'un niveau avancé et sont uniquement donnés pour indiquer les éléments de démonstration des propriétés énoncées en section 1.2.

Exercice 1.1 * Soit une suite croissante d'événements $\{A_n\}$, c'est-à-dire telle que $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$. Montrer que $P(\bigcup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} P(A_n)$. On rappelle que l'additivité des probabilités pour des événements incompatibles vaut pour une suite infinie d'événements.

Aide : considérer la suite d'événements $\{A_n \cap \overline{A}_{n-1}\}$.

Exercice 1.2 * Soit une suite décroissante d'événements $\{B_n\}$, c'est-à-dire telle que $B_1 \supseteq B_2 \supseteq \cdots \supseteq B_n \supseteq \cdots$. Montrer que $P(\bigcap_{n=1}^{\infty} B_n) = \lim_{n\to\infty} P(B_n)$.

Aide : considérer la suite $\{\overline{B}_n\}$ et admettre que le complémentaire de $\bigcap_{n=1}^{\infty} B_n$ est $\bigcup_{n=1}^{\infty} \overline{B}_n$.

Exercice 1.3 * Montrer que l'on peut écrire $F_X(+\infty) = 1$ (i.e. $\lim_{x\to+\infty} F_X(x) = 1$) et de même $F_X(-\infty) = 0$.

Aide : on utilisera le résultat de l'exercice 1.1 en considérant des événements du type $]-\infty, n]$ et $]-n, +\infty[$.

Exercice 1.4 * Montrer que $P(X = x) = F_X(x) - F_X(x^-)$.

Aide : envisager l'événement $\{x\}$ comme intersection des termes d'une suite décroissante et utiliser le résultat de l'exercice 1.2.

Exercice 1.5 Soit l'expérience aléatoire consistant à jeter un dé jusqu'à ce qu'un six apparaisse pour la première fois et soit X la v.a. «nombre de jets nécessaires». Déterminer la fonction de probabilité de X. Vérifier que la somme des probabilités sur l'ensemble des valeurs possibles est bien égale à 1. Calculer $P(1 < X \leq 3)$. Ecrire et dessiner la fonction de répartition de X.

Aide : calculer d'abord $P(X > k)$.

Exercice 1.6 Soit la fonction $f(x) = cx(1-x)$ pour $x \in [0,1]$ et 0 sinon. Pour quelle valeur de c est-ce une densité de probabilité ? Déterminer alors la fonction de répartition de cette loi et sa médiane.

Exercice 1.7 Justifier que la fonction $F(x) = 1 - e^{-\frac{x}{2}}$ pour $x > 0$ et 0 sinon, est une fonction de répartition. Déterminer les quantiles d'ordres 0,25 et 0,75 (appelés premier et troisième quartiles). Soit X une v.a. suivant cette loi, calculer $P(1 < X \leq 2)$.

Exercice 1.8 Soit X de densité $f_X(x) = 2x$ pour $x \in [0,1]$ et 0 sinon. Déterminer la fonction de répartition et la densité de $1/X$. Même question pour $\ln(1/X)$.

Exercice 1.9 Soit X de loi continue uniforme sur $[0,1]$ et $Y = -\theta \ln(1-X)$ avec $\theta > 0$. Déterminer la fonction de répartition et la densité de Y.

Chapitre 2

Espérance mathématique et moments

2.1 Introduction et définition

Dans cette section nous considérons toujours une v.a. X, soit de fonction de probabilité p_X dans le cas discret, soit de densité f_X dans le cas continu. La notion d'espérance mathématique d'une variable aléatoire correspond à la notion descriptive de moyenne pour une distribution empirique de valeurs. Nous ferons plus loin (section 2.4) une analogie entre une distribution «théorique» (une loi de probabilité) et une distribution empirique (une série d'observations numériques). Prenons pour exemple le temps de fabrication d'un produit qui connaît des variations aléatoires selon une loi supposée connue. L'espérance mathématique va indiquer quel est «en moyenne» le temps de fabrication du produit. Pour cela on effectue la somme des valeurs possibles en les affectant de poids égaux à leurs probabilités dans le cas discret, l'analogue dans le cas continu s'exprimant par une intégrale avec pondération par la densité de probabilité.

Définition 2.1 *On appelle* ***espérance mathématique*** *de X, si elle existe, la valeur notée $E(X)$ telle que :*

$$E(X) = \sum_{i=1}^{\dots} x_i \, p_X(x_i) \quad \textit{dans le cas discret,}$$

$$E(X) = \int_{-\infty}^{+\infty} x \, f_X(x) dx \quad \textit{dans le cas continu.}$$

Du point de vue du graphe de f_X (respectivement p_X) cette valeur correspond au centre de gravité de la surface sous la courbe (respectivement des bâtonnets représentant les probabilités des points). En particulier, s'il existe

un axe de symétrie, elle se situe au niveau de cet axe (par exemple l'espérance mathématique de la loi uniforme sur $[0, 360]$ est 180).

En bref, $E(X)$ sera aussi appelée *la moyenne de* X.

L'existence de $E(X)$ n'est pas garantie si f_X (respectivement p_X) converge trop lentement vers zéro à l'infini, comme dans l'exemple suivant.

Exemple 2.1 La loi de Cauchy est définie par $f_X(x) = \frac{1}{\pi(x^2+1)}$ pour $x \in \mathbb{R}$. L'espérance mathématique se calcule par

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

mais cette intégrale ne converge pas quand $x \to +\infty$ ni quand $x \to -\infty$ car la fonction à intégrer s'y comporte comme $1/x$. Plus précisément :

$$\int_a^b \frac{x}{\pi(x^2+1)} dx = \frac{1}{2\pi} \left[\ln(1+x^2)\right]_a^b = \frac{1}{2\pi} \left[\ln(1+b^2) - \ln(1+a^2)\right]$$

qui tend vers $+\infty$ quand $b \to +\infty$ et vers $-\infty$ quand $a \to -\infty$. La loi de Cauchy n'admet donc pas de moyenne (et ceci bien qu'elle soit symétrique par rapport à $x = 0$). ■

2.2 Espérance d'une fonction d'une variable aléatoire

Soit $Z = g(X)$ une v.a. fonction de la v.a. X. Pour calculer $E(Z)$ on peut d'abord déterminer sa loi (donc f_Z ou p_Z) à partir de celle de X, comme nous l'avons fait en section 1.6 . Toutefois il est possible de montrer que l'on peut directement calculer $E(Z)$ sur la loi de X, à savoir à partir de f_X ou p_X (voir les exercices pour la démonstration dans le cas continu avec une fonction dérivable monotone).

Proposition 2.1 *Soit $g(X)$ une fonction de la v.a. X. Alors :*

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x)\, f_X(x) dx \quad \textit{dans le cas continu (si l'intégrale existe),}$$

$$E(g(X)) = \sum_{i=1}^{\dots} g(x_i)\, p_X(x_i) \quad \textit{dans le cas discret (si la somme existe).}$$

On voit donc que pour le calcul de $E(g(X))$ il suffit de remplacer la valeur de x par sa valeur $g(x)$ (ou x_i par $g(x_i)$).

Exemple 2.2 Considérons X de loi uniforme sur l'intervalle $[0,1]$. Sa fonction de répartition est $F_X(x) = x$ et sa densité $f_X(x) = 1$ pour $x \in [0,1]$. Soit la fonction $Z = X^2$.

Calculons d'abord $E(X)$ en établissant la loi de Z. Celle-ci est donnée par :

$$F_Z(z) = P(Z \leq z) = P(X^2 \leq z) = P(-\sqrt{z} \leq X \leq \sqrt{z}) = P(X \leq \sqrt{z})$$

puisque $P(X \leq 0) = 0$. Donc $F_Z(z) = F_X(\sqrt{z}) = \sqrt{z}$ pour $z \in [0,1]$. Ainsi $f_Z(z) = F'_z(z) = \frac{1}{2\sqrt{z}}$ pour $z \in [0,1]$ et 0 sinon, d'où :

$$E(Z) = \int_0^1 z\,\frac{1}{2\sqrt{z}}dz = \int_0^1 \frac{\sqrt{z}}{2}dz = [\frac{z^{3/2}}{3}]_0^1 = \frac{1}{3}.$$

Calculons maintenant directement :

$$E(Z) = E(X^2) = \int_0^1 x^2.1.dx = \left[\frac{x^3}{3}\right]_0^1 = \frac{1}{3}.$$

■

Note 2.1 En étendant l'intégrale de Riemann classique à l'intégrale de Riemann-Stieltjes, on peut traiter de la même façon cas discret et cas continu. Exposons succinctement cela dans le cas où l'ensemble des valeurs possibles est borné par l'intervalle $[a,b]$. Dans la mesure où une fonction g est continue sur $[a,b]$ sauf pour un ensemble dénombrable de points, on peut définir l'intégrale de Riemann $\int_a^b g(x)dx$ de façon simple en subdivisant $[a,b]$ en n intervalles réguliers délimités par :

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b.$$

Cette intégrale est alors la limite, quand $n \to \infty$, des sommes

$$\sum_{k=1}^{n} g(x_k)\,(x_k - x_{k-1}).$$

Dans le contexte qui nous intéresse, l'intégrale de Riemann-Stieltjes relative à F_X est la limite des sommes

$$\sum_{k=1}^{n} g(x_k)\ [F_X(x_k) - F_X(x_{k-1})]\ ,$$

notée $\int_a^b g(x)\,dF_X(x)$.

Ainsi, dans le cas discret ne subsistent dans la somme que les sous-intervalles où F_X varie, c'est-à-dire contenant au moins une valeur possible x_i et, au passage à la limite, ne subsistent que ces valeurs. La limite vaut alors

$$\sum_{i=1}^{\cdots} g(x_i)\ \left[F_X(x_i) - F_X(x_i^-)\right] = \sum_{i=1}^{\cdots} g(x_i)\,p_X(x_i).$$

Dans le cas continu, par le théorème de la valeur moyenne on peut écrire $F_X(x_k) - F_X(x_{k-1}) = f_X(\xi_k)\,(x_k - x_{k-1})$ où $\xi_k \in]x_{k-1}, x_k[$, et la limite donne l'intégrale de Riemann usuelle $\int_a^b g(x)\, f_X(x)\, dx$.

Notons qu'en prenant $g(x) = 1$ sur $[a, b]$, l'intégrale $\int_c^d dF_X(x)$, où $[c, d]$ est inclus dans $[a, b]$, est égale à $F_X(d) - F_X(c)$, la probabilité associée à l'intervalle $]c, d]$, que ce soit dans le cas discret ou dans le cas continu.

2.3 Linéarité de l'opérateur $E(.)$, moments, variance

Pout toute combinaison linéaire $ag(X) + bh(X)$ de fonctions g et h de X on a :

$$E(ag(X) + bh(X)) = aE(g(X)) + bE(h(X))\,.$$

Ceci découle immédiatement de la linéarité de la sommation ou de l'intégration. Voyons cela sur le cas particulier $aX + b$ pour le cas continu :

$$\begin{aligned} E(aX + b) &= \int_{-\infty}^{+\infty} (ax + b) f_X(x) dx \\ &= a \int_{-\infty}^{+\infty} x f_X(x) dx + b \int_{-\infty}^{+\infty} f_X(x) dx = aE(X) + b\,. \end{aligned}$$

Remarquons au passage que l'on peut voir b comme une *variable aléatoire certaine* (discrète), c'est-à-dire prenant cette seule valeur avec probabilité 1 et, en cohérence avec la définition de l'espérance mathématique, écrire par convention $E(b) = b$.

On notera bien que dans le cas général $E(g(X))$ n'est pas égal à $g(E(X))$, par exemple $E(X^2) \neq (E(X))^2$.

Nous en venons maintenant à la notion de moments, lesquels sont des espérances mathématiques des puissances de X. Leur intérêt vient du fait qu'ils permettent de caractériser les distributions. Ainsi nous avons déjà vu que la moyenne (puissance 1) fournit une valeur centrale. Les puissances supérieures fournissent diverses caractéristiques de la forme de la distribution.

Définition 2.2 *On appelle* ***moment simple d'ordre*** r *de la v.a.* X*, où* r *est un entier positif, la valeur (si elle existe)* $\mu_r = E(X^r)$.

Ainsi μ_1 est la moyenne de X que l'on note plus simplement μ (ou μ_X s'il y a plusieurs v.a. à distinguer). En fait les caractéristiques de forme reposent plutôt sur les moments centrés, c'est-à-dire sur les espérances mathématiques des puissances de $X - E(X)$, ou $X - \mu$, transformation de X appelée *centrage de X.*

Définition 2.3 *On appelle* ***moment centré d'ordre*** r *de la v.a.* X*, où* r *est un entier positif, la valeur (si elle existe)* $\mu'_r = E((X-\mu)^r)$.

Pour $r = 1$ on a $E(X-\mu) = E(X) - \mu = \mu - \mu = 0$ ce qui caractérise le centrage de X. Pour $r = 2$ on a la variance de X, qui est une caractéristique de dispersion de la distribution comme en statistique descriptive et, à ce titre, mérite une attention particulière.

Définition 2.4 *On appelle* ***variance de*** X*, la valeur (si elle existe) notée* $V(X)$*, définie par :*

$$V(X) = E((X-\mu)^2).$$

On la note également plus simplement par σ^2 (éventuellement σ_X^2). La racine carrée de $V(X)$, notée naturellement σ (éventuellement σ_X), est appelée *écart-type* de X.

Les moments d'ordres supérieurs sont moins utiles et nous les mentionnons pour mémoire.

Le moment centré d'ordre 3, moyennant une standardisation pour éliminer l'effet d'échelle, fournit le *coefficient d'asymétrie :*

$$\frac{E((X-\mu)^3)}{\sigma^3}$$

dont on voit qu'il est nul en cas de symétrie (nécessairement par rapport à μ).

Du moment centré d'ordre 4 on déduit le *coefficient d'aplatissement* ou *curtose* :

$$\frac{E((X-\mu)^4)}{\sigma^4} - 3$$

qui indique, en comparaison avec la loi de Gauss, le degré de concentration autour de la moyenne (pour la loi de Gauss μ'_4 est égal à $3\sigma^4$ et ce coefficient est donc nul).

En développant $(X-\mu)^r$ et en utilisant la linéarité de l'opérateur $E(.)$, on voit que μ'_r s'exprime en fonction de $\mu_1, \mu_2, \cdots, \mu_r$. En particulier, on trouve :

$$\begin{aligned}\mu'_2 &= E((X-\mu)^2) = E(X^2 - 2\mu X + \mu^2)\\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2\\ &= E(X^2) - \mu^2.\end{aligned}$$

Ceci constitue une formule très utile pour le calcul de la variance que nous conviendrons d'appeler *formule de décentrage de la variance :*

$$V(X) = E(X^2) - \mu^2.$$

Cette formule s'apparente à celle de la statistique descriptive pour le calcul de la variance d'une série numérique (on verra dans la section suivante une analogie directe entre variance probabiliste et variance descriptive). Comme en statistique descriptive la variance ne peut être négative. De même, on a :

$$V(aX+b) = a^2 V(X).$$

Pour le voir il suffit d'appliquer la définition de la variance à la v.a. $aX+b$:

$$\begin{aligned} V(aX+b) &= E([aX+b-E(aX+b)]^2) \\ &= E([a(X-E(X))]^2) \quad \text{puisque } E(b)=b \text{ et } E(aX)=aE(X) \\ &= E(a^2[X-E(X)]^2) \qquad \text{et, encore par linéarité,} \\ V(aX+b) &= a^2 E([X-E(X)]^2) = a^2 V(X). \end{aligned}$$

Notons au passage que si X a une variance nulle c'est nécessairement une variable aléatoire certaine. En effet, pour le cas continu :

$$V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)dx$$

ne peut s'annuler puisque f_X est non négative et ne peut être nulle partout. Pour le cas discret :

$$V(X) = \sum_{i=1}^{\cdots} (x_i-\mu)^2 p_X(x_i)$$

ne peut s'annuler dès lors qu'il y a deux valeurs possibles. Inversement, si X est certaine sa variance est évidemment nulle, de sorte qu'**une variable aléatoire est certaine si et seulement si sa variance est nulle**.

Existence des moments

Si μ_r existe alors les moments d'ordres inférieurs $\mu_{r-1}, \mu_{r-2}, \cdots, \mu_1$ existent, et donc μ'_r existe. En effet la fonction x^{r-1}étant dominée par la fonction x^r au voisinage de $+\infty$ ou de $-\infty$, la convergence de l'intégrale (ou de la somme) contenant x^r entraîne celle de l'intégrale contenant x^{r-1}. Notons, pour mémoire, que la variance existe si et seulement si μ_2 existe. Par ailleurs, pour l'existence du moment d'ordre r, la convergence de $\int_{-\infty}^{+\infty} |x^r|\, f_X(x)dx = E(|X^r|)$ est une condition suffisante (ou la convergence de $\sum_{i=1}^{\cdots} |x_i^r|\, p_X(x_i)$ dans le cas discret).

2.4 Tirage aléatoire dans une population finie : distribution empirique et distribution probabiliste

La relation entre distribution empirique et distribution probabiliste suite à un tirage aléatoire permet, en particulier, de mieux appréhender les notions de moyenne et de variance d'une v.a. en les reliant aux notions correspondantes de la statistique descriptive. Considérons une population de N individus sur lesquels s'observe un certain caractère quantitatif $\mathfrak{X}$ (par exemple l'âge arrondi en années). Supposons qu'il y ait, dans cette population, r valeurs distinctes (avec $2 \leq r \leq N$) notées $x_1, x_2, \cdots, x_r$ et s'observant avec des *fréquences relatives* (fréquences[1] divisées par N) $p_1, p_2, \cdots, p_r$. La moyenne observée dans la population est donc $\sum_{i=1}^{r} x_i\, p_i$.

Considérons maintenant la v.a. X «valeur d'un individu tiré au hasard dans cette population». Par tirage au hasard, on entend que chaque individu a la même probabilité $1/N$ d'être sélectionné. De cette équiprobabilité il découle que la probabilité d'observer la valeur x_i est la fréquence relative p_i de cette valeur dans la population (voir note 1.1, deuxième paragraphe). Il y a donc identité entre la distribution empirique de $\mathfrak{X}$ dans cette population et la distribution (plus exactement la loi) de la v.a.discrète X. En particulier $E(X)$ et $V(X)$ sont identiques à la moyenne et à la variance du caractère $\mathfrak{X}$ dans la population (en prenant bien le diviseur naturel N pour le calcul de la variance de cette dernière). Pour la moyenne, la formule indiquée ci-dessus est la même que celle d'une v.a. discrète vue en section 2.1 et il en serait naturellement de même pour la variance.

2.5 Fonction génératrice des moments

La fonction génératrice des moments nous intéresse dans la mesure où elle peut faciliter le calcul des moments d'une loi. Cependant son existence - et donc son usage - sera limitée aux lois dont la densité (éventuellement la fonction de probabilité) décroît plus vite qu'une exponentielle à l'infini (voir plus loin en note 2.4 la fonction caractéristique des moments qui, elle, est toujours définie). Nous supposons ci-après qu'elle existe au moins au voisinage de 0 et que la loi admet des moments de tous ordres.

Définition 2.5 *On appelle* ***fonction génératrice des moments*** *de la v.a. X, si elle existe, la fonction :*

$$\Psi_X(t) = E(e^{tX}).$$

[1]Nous suivons l'usage anglo-saxon commode selon lequel une fréquence est un effectif et une fréquence relative est une proportion.

C'est une fonction de t par la variable t introduite dans la fonction aléatoire e^{tX}.

Proposition 2.2 *Le moment d'ordre r de la v.a. X est donné par :*

$$\mu_r = \Psi_X^{(r)}(0)$$

où $\Psi_X^{(r)}$ *est la dérivée d'ordre r de* Ψ_X. *En particulier l'espérance mathématique* (μ_1) *de X est la valeur de la dérivée première* $\Psi_X^{'}$ *pour* $t = 0$.

Note 2.2 En supposant remplies les conditions requises pour les écritures suivantes on a :

$$E(e^{tX}) = \int_{-\infty}^{+\infty} \left[\sum_{k=0}^{\infty} \frac{(tx)^k}{k!} \right] f_X(x)dx = \sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{-\infty}^{+\infty} x^k f_X(x)dx = \sum_{k=0}^{\infty} \mu_k \frac{t^k}{k!}$$

et, par identification avec le développement de $\Psi_X(t)$ en série de Taylor-Mac-Laurin, on a bien la propriété ci-dessus.

Exemple 2.3 *loi exponentielle.*

Cette loi continue, qui dépend d'un paramètre $\lambda > 0$, a pour densité (voir section 4.2.2) :

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

Calculons la fonction génératrice des moments d'une v.a. X qui suit cette loi :

$$\Psi_X(t) = E(e^{tX}) = \int_0^{+\infty} e^{tx}\, \lambda e^{-\lambda x} dx = \lambda \int_0^{+\infty} e^{(t-\lambda)x} dx\,,$$

puis en posant $u = (t - \lambda)x$ et en supposant $t < \lambda$ pour la convergence

$$\Psi_X(t) = -\frac{\lambda}{t-\lambda} \int_{-\infty}^{0} e^u du = \frac{\lambda}{\lambda - t}.$$

Les deux premières dérivées sont :

$$\Psi_X^{'}(t) = \frac{\lambda}{(\lambda - t)^2} \quad , \quad \Psi_X^{''}(t) = \frac{2\lambda}{(\lambda - t)^3} \quad ,$$

donc $\mu_1 = \Psi_X^{'}(0) = \frac{1}{\lambda}$ et $\mu_2 = \Psi_X^{''}(0) = \frac{2}{\lambda^2}$.

On obtient ainsi rapidement $E(X) = 1/\lambda$ et, par la formule de décentrage, $V(X) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$.

On pourrait obtenir aussi aisément les moments d'ordres supérieurs. ■

Note 2.3 Si l'on sait développer $\Psi_X(t)$en série entière, $\Psi_X(t) = \sum_{k=0}^{\infty} a_k t^k$, on accède directement aux moments. Le moment d'ordre k est en effet le coefficient du terme en t^k, multiplié par $k!$ (voir note 2.2 ci-dessus). Par exemple pour la loi exponentielle on a :

$$\Psi_X(t) = \frac{1}{(1-\frac{t}{\lambda})} = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \cdots + \frac{t^k}{\lambda^k} + \cdots$$

et μ_k est donc égal à $k!/\lambda^k$, comme on l'a vu pour $k = 1$ et $k = 2$.

Exemple 2.4 *loi géométrique.*

Cette loi discrète, qui dépend d'un paramètre $p \in [0, 1]$, a pour fonction de probabilité (voir section 4.1.4) :

$$p_X(x) = p(1-p)^x \qquad \text{pour } x = 0, 1, 2, \cdots .$$

On a alors :

$$\Psi_X(t) = \sum_{x=0}^{\infty} e^{tx} p(1-p)^x = p \sum_{x=0}^{\infty} \left[(1-p)e^t\right]^x$$

qui est définie pour $(1-p)e^t < 1$ ou $t < \log(\frac{1}{1-p})$. Ainsi :

$$\Psi_X(t) = \frac{p}{1-(1-p)e^t}$$

$$\Psi_X^{'}(t) = \frac{p(1-p)e^t}{\left[1-(1-p)e^t\right]^2}$$

$$\begin{aligned}\Psi_X^{''}(t) &= \frac{p(1-p)e^t}{\left[1-(1-p)e^t\right]^2} - 2\frac{p(1-p)e^t[-(1-p)e^t]}{\left[1-(1-p)e^t\right]^3} \\ &= \frac{p(1-p)e^t[1-(1-p)e^t + 2(1-p)e^t]}{\left[1-(1-p)e^t\right]^3} \\ &= \frac{p(1-p)e^t[1+(1-p)e^t]}{\left[1-(1-p)e^t\right]^3}\end{aligned}$$

d'où :

$$\mu_1 = \frac{p(1-p)}{p^2} = \frac{1-p}{p}$$

$$\mu_2 = \frac{p(1-p)(2-p)}{p^3} = \frac{(1-p)(2-p)}{p^2}$$

et

$$V(X) = \frac{(1-p)(2-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}\,.$$

■

Note 2.4 *Fonction caractéristique des moments* Φ_X

Elle est définie via une extension à des variables aléatoires à valeurs dans l'ensemble des nombres complexes $\mathbb{C}$ par :

$$\Phi_X(t) = E(e^{itX}) = E(\cos(tX)) + iE(\sin(tX)).$$

Puisque f_X est intégrable sur tout $\mathbb{R}$ et que $\left|e^{itx}\right| \leq 1$ pour tout x, $\Phi_X(t)$ est définie pour tout t, quelle que soit la loi (le même type d'argument valant pour une loi discrète avec p_X). En fait, la fonction caractéristique des moments permet de définir parfaitement une loi de probabilité, et ceci de façon duale avec la fonction de répartition. Moyennant les conditions nécessaires de dérivabilité on a $\mu_k = i^{-k}\Phi_X^{(k)}(0)$. On recourra éventuellement à la fonction caractéristique lorsque la fonction génératrice n'existera pas au voisinage de 0. Quand cette dernière existe on en déduit immédiatement $\Phi_X(t) = \Psi_X(it)$.

2.6 Formules d'approximation de l'espérance et de la variance d'une fonction d'une v.a.

Ces formules sont utiles car on se heurte souvent à un problème d'intégration (ou de sommation) pour le calcul de $E(g(X))$ ou de $V(g(X))$. Nous adoptons les mêmes notations qu'en section 2.2 et supposons que g est dérivable deux fois au voisinage de $\mu = E(X)$. En développant $g(x)$ en série de Taylor au voisinage de μ :

$$g(x) = g(\mu) + (x-\mu)g'(\mu) + \frac{(x-\mu)^2}{2}g''(\mu) + o((x-\mu)^2)$$

et en négligeant le terme[2] $o((x-\mu)^2)$ on obtient :

$$\begin{aligned} E(g(X)) &\simeq g(\mu) + g'(\mu)E(X-\mu) + g''(\mu)\frac{E((X-\mu)^2)}{2} \\ &\simeq g(\mu) + \frac{1}{2}g''(\mu)\,V(X)\,, \end{aligned}$$

puis :

$$\begin{aligned} g(x) - E(g(X)) &\simeq (x-\mu)\,g'(\mu) + \frac{1}{2}[(x-\mu)^2 - V(X)]g''(\mu) \\ [g(x) - E(g(X))]^2 &\simeq (x-\mu)^2[g'(\mu)]^2\,, \end{aligned}$$

d'où :

$$V(g(X)) \simeq V(X)[g'(\mu)]^2\,.$$

Le même type d'approximation peut être obtenu pour les fonctions de plusieurs variables définies au chapitre suivant.

[2]La notation $o(u)$ est utilisée pour désigner une fonction telle que $\frac{o(u)}{u} \to 0$ quand $u \to 0$, c'est-à-dire qu'elle devient négligeable par rapport à u quand u est petit.

2.7 Exercices

Exercice 2.1 Soit la v.a. discrète X prenant pour valeurs 0, 1, 2 avec les probabilités respectives 0,7 ; 0,2 ; 0,1. Soit $Y = X^2 - 1$. Calculer $E(Y)$.

Exercice 2.2 Soit la v.a. X continue, de fonction de répartition F_X et de densité f_X, et soit la fonction $Z = g(X)$ où g est une fonction strictement croissante (continûment) dérivable. En appliquant la méthode décrite en section 1.6, déterminer la fonction de répartition de Z et en déduire que sa densité est $f_Z(z) = f_X(g^{-1}(z))\,|(g^{-1}(z))'\,|$. On montrera que ceci reste vrai pour une fonction strictement décroissante.

Établir la propriété énoncée dans la proposition 2.1, i.e. $\int_{-\infty}^{+\infty} g(x)\, f_X(x)dx = E(Z)$ (aide : utiliser le changement de variable $x = g^{-1}(z)$ dans l'intégrale).

Exercice 2.3 Soit la loi (dite exponentielle double ou loi de Laplace) de densité :

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

Montrer que sa fonction génératrice des moments est $\Psi(t) = (1 - t^2)^{-1}$. En déduire sa variance et son moment d'ordre 4. Calculer son coefficient d'aplatissement (défini en section 2.3).

Exercice 2.4 Soit la loi de Pareto (voir section 4.2.6) de paramètres strictement positifs a et θ, dont la fonction de densité est :

$$f(x) = \begin{cases} \dfrac{\theta}{a}\left(\dfrac{a}{x}\right)^{\theta+1} & \text{si } x \geq a \\ 0 & \text{si } x < a \end{cases}\,.$$

1. Calculer la moyenne et la variance de cette loi. Quand ces moments existent-ils ? Généraliser à l'existence d'un moment d'ordre quelconque.
2. Montrer que sa fonction génératrice des moments n'existe pas.

Exercice 2.5 Soit la v.a. X de densité $f(x) = 3x^2$ si $x \in [0, 1]$ et 0 sinon.
1. Calculer $E(1/X)$.
2. Déterminer la fonction de répartition de $Y = 1/X$ et en déduire sa densité. Calculer $E(Y)$ et vérifier ainsi le résultat obtenu au point précédent.

Les notions de ce chapitre seront largement illustrées au cours du chapitre 4 sur les lois usuelles.

Chapitre 3

Couples et n-uplets de variables aléatoires

3.1 Introduction

Dans ce chapitre nous ne développerons l'étude simultanée de plusieurs v.a. que de façon restreinte en ne présentant que ce qui est nécessaire pour préparer l'approche statistique ultérieure.

Dans un premier volet (sections 3.2 à 3.5) nous étudierons les couples de v.a. en mettant en évidence la façon de formaliser la relation entre deux quantités aléatoires. Nous introduirons notamment les notions de covariance et de corrélation qui répondent, dans un cadre probabiliste, aux notions du même nom de la statistique descriptive. L'étude des relations deux à deux entre plusieurs variables aléatoires nous conduira, en section 3.8, à introduire la notation matricielle, en particulier avec la matrice des variances-covariances sur laquelle repose essentiellement la statistique «multivariée». En ce sens l'étude des couples de variables aléatoires est un point de départ suffisant pour aborder la théorie multivariée.

Dans un deuxième temps (sections 3.6 et 3.7) nous porterons notre attention sur une suite de n v.a., non pas pour étudier le jeu de leurs relations, mais comme prélude aux propriétés des échantillons aléatoires qui sont les objets principaux de la statistique mathématique. En effet nous verrons qu'un échantillon aléatoire de taille n se définit comme une suite de n v.a. indépendantes et de même loi, ce qui correspond à l'observation répétée du même phénomène quantitatif. La relation entre ces observations n'est pas pertinente vu leur caractère d'indépendance.

3.2 Couples de v.a.

Nous nous intéressons donc à l'étude de deux entités numériques a priori aléatoires, par exemple le poids et la taille d'un individu (cas continu), le nombre d'enfants et le nombre de pièces du logement d'un ménage (cas discret). Nous nous contenterons, comme nous l'avons fait pour une seule v.a., d'une définition informelle.

Un couple de v.a. peut être vu comme un ensemble Ω de valeurs de $\mathbb{R}^2$ auquel on associe une mesure de probabilité. Comme pour le cas d'une v.a. simple (voir section 1.1) la mesure de probabilité est une fonction portant sur l'ensemble des événements, lesquels sont des parties de $\mathbb{R}^2$. La fonction de répartition conjointe sera l'instrument fondamental pour donner la probabilité d'une région quelconque du plan (quoique dans le cas discret on préférera, en pratique, recourir à la fonction de probabilité conjointe).

Dans ce qui suit nous désignerons de façon générale par (X, Y) le couple de variables aléatoires. Par simplicité, nous ne considérons que des couples où les deux variables sont de même nature, discrètes ou continues, et exclurons le cas mixte.

Définition 3.1 *Soit (X, Y) un couple de v.a., on appelle* ***fonction de répartition conjointe*** *de (X, Y), que l'on note $F_{X,Y}$, la fonction définie sur $\mathbb{R}^2$ par :*

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

Dans ces notations, précisons que l'événement $(X \leq x, Y \leq y)$ peut se lire $(X \leq x) \cap (Y \leq y)$, c'est-à-dire qu'à la fois X soit inférieur ou égal à x et Y soit inférieur ou égal à y.

Note 3.1 En principe la fonction de répartition conjointe suffit à calculer la probabilité de tout événement car les seules parties de $\mathbb{R}^2$ probabilisées sont celles générées par les unions, intersections et compléments de parties du type $(X \leq x, Y \leq y)$, formant la tribu borélienne de $\mathbb{R}^2$ (voir l'analogie avec une v.a. simple dans la note 1.2).

Définition 3.2 (cas discret) *Soit (X, Y) un couple de v.a. discrètes pouvant prendre les couples de valeurs $\{(x_i, y_j); i = 1, 2, \ldots ; j = 1, 2, \ldots\}$. On appelle* ***fonction de probabilité conjointe*** *la fonction, notée $p_{X,Y}$, qui donne les probabilités associées à ces couples de valeurs, soit, pour tout i et tout j :*

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j).$$

Définition 3.3 *Soit (X, Y) un couple de v.a. continues, on appelle* ***fonction de densité de probabilité conjointe*** *la fonction non négative sur $\mathbb{R}^2$ notée*

$f_{X,Y}$ telle que :

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(u,v)dudv\,.$$

Par convention, lorsque l'on parlera d'un couple de v.a. continues, on supposera l'existence de cette fonction.

Si l'on s'intéresse à un événement sur X quelle que soit la valeur prise par Y, on retombe sur la loi de la v.a. X qui, dans le contexte du couple, est appelée *loi marginale de X*. On peut faire le lien avec la fonction de répartition conjointe en écrivant :

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(X \leq x, Y \in \mathbb{R}) \\ &= \lim_{y\to+\infty} P(X \leq x, Y \leq y) \\ &= F_{X,Y}(x,+\infty) \end{aligned}$$

De même $F_Y(y) = F_{X,Y}(+\infty, y)$.

Dans le cas discret il est clair que la fonction de probabilité marginale de X, par exemple, peut s'obtenir en sommant la fonction de probabilité conjointe sur toutes les valeurs possibles de Y, i.e. :

$$p_X(x_i) = \sum_{j=1}^{\cdots} p_{X,Y}(x_i, y_j)\,.$$

Pour le cas continu on admettra la relation du même type portant sur les densités :

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)dy.$$

On peut définir encore des *lois conditionnelles* pour l'une des variables, l'autre étant fixée à telle ou telle valeur. Nous illustrons ceci d'abord dans le cas discret qui est plus simple. Ainsi, reprenant l'exemple introductif où X est le nombre de pièces du logement d'un ménage pris au hasard et Y est le nombre d'enfants de ce ménage, nous pouvons considérer par exemple la loi de X *sachant* (Y=2). Dans l'analogie entre probabilités et fréquences relatives (voir section 2.4), cela équivaut à définir la distribution du nombre de pièces **parmi** les ménages ayant deux enfants. Plus généralement, on définira la fonction de probabilité conditionnelle de X sachant $(Y = y_j)$ en appliquant la règle des probabilités conditionnelles $P(A|B) = P(A \cap B)/P(B)$, soit, avec des notations parlant d'elles-mêmes :

$$p_{X|Y=y_j}(x_i) = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)} \quad , \quad i = 1, 2, \ldots\ .$$

On peut évidemment définir de façon similaire $p_{Y|X=x_i}(y_j)$.

Pour le cas continu, les choses sont plus compliquées car, comme nous l'avons vu, la probabilité que Y, par exemple, prenne une valeur donnée est nulle. Il n'empêche, même si cela peut paraître paradoxal au premier abord, que l'on peut définir une loi de X *sachant* $(Y = y)$, dès lors toutefois qu'en y on ait $f_Y(y) > 0$. On voit l'intérêt de ceci dans le cas de l'exemple introductif où X serait le poids d'un individu et Y sa taille (en cm). La loi de X sachant $(Y = 170)$ serait en quelque sorte, par analogie avec les fréquences relatives, la distribution des poids **parmi** les personnes mesurant 170 cm. On peut établir la fonction de répartition conditionnelle par un raisonnement limite. La probabilité de $(X \leq x)$ sachant que $(y - \frac{h}{2} \leq Y \leq y + \frac{h}{2})$ se calcule par :

$$\begin{aligned}
&\frac{P(X \leq x\,,\, y - \frac{h}{2} \leq Y \leq y + \frac{h}{2})}{P(y - \frac{h}{2} \leq Y \leq y + \frac{h}{2})} \\
&= \frac{P(X \leq x\,,\, Y \leq y + \frac{h}{2}) - P(X \leq x\,,\, Y \leq y - \frac{h}{2})}{P(y - \frac{h}{2} \leq Y \leq y + \frac{h}{2})} \\
&= \frac{F_{X,Y}(x\,,\, y + \frac{h}{2}) - F_{X,Y}(x\,,\, y - \frac{h}{2})}{F_Y(y + \frac{h}{2}) - F_Y(y - \frac{h}{2})}.
\end{aligned}$$

Pour le numérateur, on a utilisé le fait que :

$$(X \leq x, Y \leq y + \frac{h}{2}) = (X \leq x, y - \frac{h}{2} \leq Y \leq y + \frac{h}{2}) \cup (X \leq x, Y \leq y - \frac{h}{2})$$

et que ces deux derniers événements sont incompatibles. En faisant tendre h vers 0, on obtient la fonction de répartition conditionnelle de X sachant $(Y = y)$, soit, moyennant une division du numérateur comme du dénominateur par h :

$$F_{X|Y=y}(x) = \lim_{h \to 0} \frac{\left[F_{X,Y}(x\,,\, y + \frac{h}{2}) - F_{X,Y}(x\,,\, y - \frac{h}{2})\right]/h}{\left[F_Y(y + \frac{h}{2}) - F_Y(y - \frac{h}{2})\right]/h}$$

où le dénominateur tend vers la densité marginale de Y en y (voir section 1.4) et le numérateur tend vers la dérivée partielle[1] de $F_{X,Y}$ par rapport à y, au point (x, y). Cette dernière étant égale à $\int_{-\infty}^{x} f_{X,Y}(u, y)du$, on a finalement :

$$F_{X|Y=y}(x) = \frac{\int_{-\infty}^{x} f_{X,Y}(u, y)du}{f_Y(y)}\,.$$

Par dérivation par rapport à x, on obtient la densité conditionnelle :

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

dont l'expression rappelle celle de la fonction de probabilité conditionnelle du cas discret.

[1] Tout comme en section 1.4 il a été dit que F_X est dérivable partout sauf peut-être sur un ensemble dénombrable de points, $F_{X,Y}$ sera dérivable par rapport à x et à y partout sauf éventuellement sur une partie de $\mathbb{R}^2$ de probabilité nulle.

3.3 Indépendance de deux variables aléatoires

L'indépendance d'une v.a. X d'une part et d'une v.a. Y d'autre part se rapporte aux occurrences simultanées d'événements sur X et d'événements sur Y. Nous devons donc partir du couple (X, Y).

Définition 3.4 *Deux v.a. X et Y sont dites **indépendantes** si, étant donné deux événements quelconques $(X \in A)$ et $(Y \in B)$, on a :*

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B).$$

Proposition 3.1 *X et Y sont indépendantes si et seulement si :*

$$\textit{pour tout } (x, y) \in \mathbb{R}^2, \quad F_{X,Y}(x, y) = F_X(x)\, F_Y(y).$$

Le fait que l'indépendance entraîne que la fonction de répartition conjointe soit le produit des deux fonctions de répartition marginales est évident en considérant les événements particuliers de la forme $(X \leq x)$ et $(Y \leq y)$. Le fait que cette condition soit également suffisante pour assurer l'indépendance tient au caractère générateur des événements du type $(X \leq x, Y \leq y)$ pour l'ensemble des événements envisagés dans $\mathbb{R}^2$. Les deux propositions suivantes, distinguant cas discret et cas continu, seront particulièrement utiles (on conserve les mêmes notations que précédemment).

Proposition 3.2 *Deux v.a. discrètes X et Y sont indépendantes si et seulement si, pour tout $i = 1, 2, \ldots$ et tout $j = 1, 2, \ldots$,*

$$p_{X,Y}(x_i, y_j) = p_X(x_i)\, p_Y(y_j)\,.$$

Proposition 3.3 *Deux v.a. continues X et Y sont indépendantes si et seulement si, pour tout $(x, y) \in \mathbb{R}^2$,*

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

Proposition 3.4 *Si X et Y sont indépendantes, alors pour toutes fonctions g et h, les v.a. $g(X)$ et $h(Y)$ sont également indépendantes.*

Ce résultat est immédiat dans la mesure où tout événement sur $g(X)$ peut s'exprimer comme un événement sur X et de même pour $h(Y)$ sur Y.

Note 3.2 Il va de soi que ces fonctions doivent être mesurables (voir note 1.1).

3.4 Espérance mathématique, covariance, corrélation

Pour le couple de v.a. (X, Y) nous connaissons déjà $E(X)$ et $E(Y)$, moyennes respectives des lois marginales de X et de Y. De façon semblable à ce qui a été fait en section 2.2, nous pouvons aussi définir la notion d'espérance mathématique d'une fonction $g(X, Y)$ du couple. En particulier, dans l'approche statistique on utilisera abondamment la fonction somme $X + Y$, que nous étudierons plus spécialement dans la section suivante comme une application de la section présente.

Étant donné $g(X, Y)$ une v.a. à valeurs dans $\mathbb{R}$, selon le même principe qu'en section 2.2, nous pouvons directement déterminer son espérance mathématique en considérant les images par g de toutes les valeurs possibles du couple (X, Y) et en les pondérant par les probabilités (ou densités de probabilité pour le cas continu) correspondantes. D'où, moyennant l'existence des doubles sommes ou des intégrales doubles :

$$E(g(X,Y)) = \sum_{i=1}^{\cdots} \sum_{j=1}^{\cdots} g(x_i, y_j) p_{X,Y}(x_i, y_j) \quad \text{dans le cas discret,}$$

$$E(g(X,Y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f_{X,Y}(x,y) dxdy \quad \text{dans le cas continu.}$$

On pourrait également établir la loi de cette nouvelle v.a. $Z = g(X, Y)$ et calculer sa moyenne $E(Z)$. Mais nous ne traiterons pas ici la façon d'obtenir la loi d'une fonction de deux variables aléatoires.

Proposition 3.5 (linéarité de l'espérance mathématique) *Pour la fonction $aX + bY$ on a :*

$$E(aX + bY) = aE(X) + bE(Y).$$

Ceci est une extension du résultat donné en début de section 2.3 et découle également de la linéarité des sommations et intégrations doubles. Cette propriété reste évidemment valable si l'on substitue à X une v.a. $g(X)$ et à Y une v.a. $h(Y)$, par exemple :

$$E(2X^2 + 3Y^2) = 2E(X^2) + 3E(Y^2).$$

Les notions de moments simples ou centrés vues en section 2.3 s'étendent au cas du couple. En bref, on définit le *moment simple croisé d'ordres* (p, q) par $E(X^p Y^q)$ et le moment *centré* correspondant par $E([X - E(X)]^p [Y - E(Y)]^q)$. Seul le cas $p = 1$ et $q = 1$ mérite notre intérêt, conduisant notamment aux notions clés de covariance et corrélation entre deux variables aléatoires.

Définition 3.5 *On appelle* ***covariance de*** X **et de** Y, *que l'on note* $cov(X,Y)$, *le nombre (s'il existe) :*

$$cov(X,Y) = E([X - E(X)][Y - E(Y)]).$$

On remarquera d'emblée que c'est une notion symétrique en X et Y, i.e. $cov(X,Y) = cov(Y,X)$.

Proposition 3.6 (formule de décentrage de la covariance)

$$cov(X,Y) = E(XY) - E(X)E(Y).$$

En effet :

$$\begin{aligned} cov(X,Y) &= E([X - E(X)][Y - E(Y)]) \\ &= E(XY - E(X)Y - X E(Y) + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y)\,. \end{aligned}$$

Cette formule est à rapprocher de la formule de décentrage de la variance vue en section 2.3. D'ailleurs, on notera que $cov(X,X) = E([X-E(X)]^2) = V(X)$.

Proposition 3.7 *Si X et Y sont indépendantes alors* $cov(X,Y) = 0$.

En effet il suffit de vérifier qu'en cas d'indépendance $E(XY) = E(X)E(Y)$. Faisons-le dans le cas continu, par exemple, en rappelant que l'indépendance implique $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

$$\begin{aligned} E(XY) &= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy f_X(x) f_Y(y) dx dy \\ &= \int_{-\infty}^{+\infty} y f_Y(y) \left[\int_{-\infty}^{+\infty} x f_X(x) dx\right] dy \\ &= E(X)\int_{-\infty}^{+\infty} y f_Y(y) dy \\ &= E(X)E(Y). \end{aligned}$$

Notons bien que **deux v.a. peuvent avoir une covariance nulle sans pour autant être indépendantes**. Montrons-le sur un exemple artificiel**.**

Exemple 3.1 Soient X et Y deux variables aléatoires discrètes, chacune pouvant prendre les valeurs 0 ,1 ou 2. Les probabilités conjointes sont données à l'intérieur du tableau croisé ci-dessous, les marges représentant les probabilités marginales.

X \ Y	0	1	2	
0	0	4/9	0	4/9
1	2/9	0	2/9	4/9
2	0	1/9	0	1/9
	2/9	5/9	2/9	1

On a :

$$E(X) = \frac{4}{9} + \frac{2}{9} = \frac{2}{3}\,, \quad E(Y) = \frac{5}{9} + 2 \times \frac{2}{9} = 1,$$
$$E(XY) = 1 \times 2 \times \frac{2}{9} + 2 \times 1 \times \frac{1}{9} = \frac{2}{3}$$

d'où $cov(X,Y) = E(XY) - E(X)E(Y) = 0$. Or X et Y ne sont pas indépendantes puisque, par exemple, $P(X = 0, Y = 0)$ est nul alors que $P(X=0)P(Y=0) = \frac{2}{9} \times \frac{4}{9} \neq 0$. ■

Propriétés de la covariance

1. $cov(aX + b, cY + d) = ac\, cov(X,Y)$
2. $cov(X + Y, Z) = cov(X,Z) + cov(Y,Z)$

Pour montrer le point 1, appliquons la définition de la covariance aux v.a. $aX + b$ et $cY + d$:

$$\begin{aligned} cov(aX + b, cY + d) &= E([aX + b - E(aX + b)][cY + d - E(cY + d]) \\ &= E([aX + b - aE(X) - b][cY + d - cE(Y) - d]) \\ &= E(a[X - E(X)]\, c[Y - E(Y)]) = ac\, cov(X,Y). \end{aligned}$$

On voit que les constantes disparaissent en raison des centrages effectués par la covariance. Le point 2 se démontre en développant de façon analogue.

Définition 3.6 *On appelle (coefficient de)* ***corrélation linéaire*** *de X et de Y, que l'on note $corr(X,Y)$, le nombre (s'il existe) :*

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X\ \sigma_Y}$$

où σ_X est l'écart-type de X et σ_Y celui de Y.

Cette formule s'apparente à celle de la statistique descriptive pour le calcul de la corrélation linéaire sur une série d'observations couplées. Dans le cas particulier du tirage aléatoire d'un individu dans une population vu en section 2.5, la corrélation descriptive devient la corrélation probabiliste. Supposons que l'on étudie la population des ménages résidant dans une ville donnée, les variables considérées étant $\mathcal{X}$ le nombre de pièces du logement et $\mathcal{Y}$ le nombre

d'enfants. Soit f_{ij} la fréquence relative, dans la population, du couple de valeurs (x_i, y_j). Le coefficient de corrélation linéaire descriptif (ou empirique) du lien entre nombre de pièces et nombre d'enfants dans cette population est :

$$\frac{\sum_{i=1}^{\dots}\sum_{j=1}^{\dots}(x_i-\overline{x})(y_j-\overline{y})f_{ij}}{\sqrt{\sum_{i=1}^{\dots}(x_i-\overline{x})^2 f_i}\ \sqrt{\sum_{j=1}^{\dots}(y_j-\overline{y})^2 f_j}}$$

où $\overline{x} = \sum_{i=1}^{\dots} x_i f_i$ est le nombre moyen de pièces pour l'ensemble des ménages avec f_i égal à la fréquence du nombre de pièces x_i et, de la même façon avec f_j, $\overline{y}$ est le nombre moyen d'enfants. En tirant un ménage au hasard, on induit un couple de v.a. (X, Y) pour lequel la probabilité associée au couple de valeurs (x_i, y_j) devient f_{ij}. Le numérateur de la formule ci-dessus est alors $cov(X, Y)$ et le dénominateur $\sigma_X\ \sigma_Y$.

Nous énonçons ci-après quelques propriétés de la corrélation linéaire, identiques à celles de la statistique descriptive.

Propriétés de la corrélation linéaire

1. $corr(X, Y) = corr(Y, X)$
2. $corr(aX + b, cY + d) = corr(X, Y)$

La propriété 1 est évidente. La propriété 2 résulte du fait que $\sigma_{aX+b} = a\ \sigma_X$ puisque $V(aX+b) = a^2V(X)$ (voir section 2.3) et $\sigma_{cY+d} = c\,\sigma_Y$. Le produit ac obtenu dans la covariance disparaît donc en divisant par le produit des écarts-types. Cette propriété indique que la corrélation linéaire entre deux v.a. est invariante dans un changement d'échelle (et même un changement d'origine comme pour le passage d'une température en Celsius à une température en Fahrenheit), ce qui semble raisonnable pour une mesure de lien entre deux entités numériques.

Proposition 3.8 *Quelle que soit la loi conjointe du couple (X, Y) on a :*

$$-1 \leq corr(X, Y) \leq 1\ .$$

Démonstration : considérons la v.a. $X+\lambda Y$ où λ est un nombre réel quelconque. Par définition de la variance :

$$\begin{aligned}
V(X+\lambda Y) &= E([X+\lambda Y - E(X+\lambda Y)]^2) \\
&= E([X-E(X)+\lambda(Y-E(Y))]^2) \\
&= E([X-E(X)]^2 + 2\lambda[X-E(X)][Y-E(Y)] + \lambda^2[Y-E(Y)]^2) \\
&= E([X-E(X)]^2) + 2\lambda E([X-E(X)][Y-E(Y)]) + \lambda^2 E([Y-E(Y)]^2) \\
&= V(X) + 2\lambda cov(X,Y) + \lambda^2 V(Y)\ .
\end{aligned}$$

Or $V(X + \lambda Y) \geq 0$ **quel que soit** λ. Cela implique, pour le polynôme du deuxième degré en λ ci-dessus, que le déterminant $[cov(X,Y)]^2 - V(X)V(Y)]$ est négatif ou nul, et donc :

$$[cov(X,Y)]^2 \leq V(X)V(Y)$$
$$\text{et} \qquad [corr(X,Y)]^2 = \frac{[cov(X,Y)]^2}{V(X)V(Y)} \leq 1 \,.$$

□

Notons que si $[corr(X,Y)]^2 = 1$ alors le déterminant est nul et, pour la racine double λ_0 du polynôme, on a $V(X + \lambda_0 Y) = 0$, c'est-à-dire (voir section 2.3) que $X + \lambda_0 Y$ est une v.a. certaine. En d'autres termes, il existe une dépendance linéaire parfaite entre X et Y. Ceci est à rapprocher du fait qu'en statistique descriptive une corrélation égale à ± 1 équivaut à un alignement parfait des points représentant les couples de valeurs. Ce résultat justifie aussi l'appellation de corrélation *linéaire*.

La corrélation s'annule si et seulement si la covariance s'annule et, donc, **une corrélation nulle n'implique pas l'indépendance.**

3.5 Somme de deux v.a.

Etudions la fonction particulière $X+Y$ issue du couple (X,Y) en vue d'une généralisation dans la section suivante à une somme de n v.a. qui, comme il a été dit plus haut, sera un objet essentiel de la statistique. Nous savons déjà que, par la linéarité,

$$E(X+Y) = E(X) + E(Y)\,.$$

Proposition 3.9 *On a :*

$$V(X+Y) = V(X) + V(Y) + 2cov(X,Y)\,.$$

Si X et Y sont indépendantes, alors :

$$V(X+Y) = V(X) + V(Y).$$

La première équation est le cas particulier $\lambda = 1$ du développement de $V(X + \lambda Y)$ dans la démonstration de la proposition 3.8. La deuxième est évidente puisque l'indépendance implique une covariance nulle.

Proposition 3.10 (fonctions génératrices des moments) *Si X et Y sont indépendantes, alors :*

$$\Psi_{X+Y}(t) = \Psi_X(t)\Psi_Y(t).$$

En effet $\Psi_{X+Y}(t) = E(e^{(X+Y)t}) = E(e^{Xt}e^{Yt}) = E(e^{Xt})E(e^{Yt})$ puisque si X et Y sont indépendantes, alors les fonctions e^{Xt} et e^{Yt} sont également indépendantes (voir proposition 3.4).

Note 3.3 Il en va évidemment de même pour la fonction caractéristique des moments.

3.6 Les n-uplets de v.a. ; somme de n v.a.

Nous considérons maintenant la généralisation du couple (ou 2-uplet) à un n-uplet, c'est-à-dire à un **vecteur aléatoire,** prenant ses valeurs dans $\mathbb{R}^n$ et que nous noterons $(X_1, X_2, \cdots, X_n)$. Comme il a été dit en introduction notre intérêt va se porter essentiellement sur le cas particulier d'un échantillon aléatoire et c'est pourquoi nous ne nous attardons pas sur la loi conjointe du n-uplet. Disons brièvement que la notion de fonction de répartition conjointe se généralise naturellement selon :

$$F_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n) \ .$$

Les notions de lois marginales et lois conditionnelles se généralisent de la même façon. On peut ainsi définir des lois marginales pour tout sous-ensemble de composantes du vecteur aléatoire $(X_1, X_2, \cdots, X_n)$. On peut définir des lois conditionnelles d'un sous-ensemble sachant les valeurs d'un autre sous-ensemble. La notion de covariance (respectivement corrélation) se généralise en matrice des variances-covariances (respectivement des corrélations) des composantes prises deux à deux (voir section 3.8). La notion d'indépendance dans un couple se généralise à la notion d'indépendance mutuelle des n composantes selon laquelle les événements portant sur tous sous-ensembles sont indépendants. Par simplification nous omettrons l'adjectif «mutuelle» qui sera implicite.

Nous limitons désormais notre étude des n-uplets au cas où les n composantes ont **la même loi de probabilité** marginale et sont **indépendantes**. On peut considérer alors le n-uplet comme n observations successives d'un même phénomène aléatoire, ces observations étant indépendantes les unes des autres (au sens où le fait que, par exemple, la première observation donne telle valeur n'influe en rien sur le résultat de la deuxième observation). On voit poindre ici de façon évidente l'approche statistique, de tels n-uplets constituant précisément ce que nous appellerons plus loin des **échantillons aléatoires**. Pour l'heure nous donnons des résultats sur les fonctions **somme** et **moyenne** des n composantes (conduisant plus loin à l'étude du comportement de la somme ou de la moyenne d'un échantillon aléatoire).

Proposition 3.11 (proposition fondamentale) *Soit $X_1, X_2, \cdots, X_n$ une suite de v.a. indépendantes et suivant une même loi de probabilité de moyenne μ et de variance σ^2. On a, pour la somme $S_n = X_1 + X_2 + \cdots + X_n$:*

$$E(S_n) = n\mu \qquad V(S_n) = n\sigma^2$$

et pour la moyenne $\overline{X}_n = S_n/n$ *:*

$$E(\overline{X}_n) = \mu \qquad V(\overline{X}_n) = \frac{\sigma^2}{n} \ .$$

Pour S_n ces résultats sont la généralisation de proche en proche, d'une part de la propriété de linéarité de l'espérance mathématique, à savoir $E(S_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$, d'autre part de l'additivité des variances pour des variables indépendantes, à savoir $V(S_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$.

Pour la moyenne on a :

$$E(\overline{X}_n) = E\left(\frac{S_n}{n}\right) = \frac{1}{n}E(S_n) = \frac{n\mu}{n} = \mu$$
$$V(\overline{X}_n) = V\left(\frac{S_n}{n}\right) = \frac{1}{n^2}V(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \ .$$

Notons, pour mémoire, que l'écart-type de $\overline{X}_n$ est $\dfrac{\sigma}{\sqrt{n}}$.

Des v.a. indépendantes et de même loi sont couramment notées en bref **variables aléatoires i.i.d**., abréviation de «indépendantes et identiquement distribuées». Nous adopterons dorénavant cette notation.

Proposition 3.12 ***(fonction génératrice d'une somme de v.a. i.i.d.)*** *Soit* $\Psi(t)$ *la fonction génératrice des moments de la loi commune aux* n *v.a. i.i.d., alors* $\Psi_{S_n}(t) = [\Psi(t)]^n$.

Ceci est une extension évidente de proche en proche de la proposition 3.10 sur la somme de deux v.a. .

3.7 Sondage aléatoire dans une population et v.a. i.i.d.

Ayant franchi un pas vers l'idée d'observations répétées, on peut se poser la question de savoir comment se traduit en termes probabilistes l'expérience consistant à effectuer non plus un seul tirage aléatoire comme vu en section 2.4, mais n tirages successifs d'individus dans une population. S'agissant d'une population bien réelle de N individus, ce contexte expérimental est celui du sondage.

Un *sondage aléatoire simple* (par distinction vis-à-vis de plans de sondage plus complexes) consiste à sélectionner un premier individu avec équiprobabilité de tous les N individus, puis un deuxième individu avec équiprobabilité des $N-1$ individus restants et ainsi de suite pour les $N-2$ individus restants, etc. jusqu'à sélectionner n individus. Pour une variable quantitative d'intérêt sur les individus, notons X_1 l'observation aléatoire du premier tirage, X_2 celle du deuxième tirage, ..., X_n celle du n-ième tirage. Constatons que, dans ce schéma, il n'y a pas indépendance des v.a. $X_1, X_2, \cdots, X_n$. Par exemple, au deuxième tirage, les valeurs possibles (recevant la probabilité $\frac{1}{N-1}$) sont sujettes au résultat du premier tirage. Le sondage n'est donc pas une situation de v.a. i.i.d., ce qui complique tant soit peu les choses et explique que la théorie des sondages occupe une place à part dans la statistique mathématique. Une façon de contourner le problème de la dépendance consisterait à effectuer un sondage *avec remise (* par opposition au sondage usuel précédent que l'on qualifie de *sans remise*), c'est-à-dire à réintégrer dans la population, à chaque tirage, l'individu tiré. Mais ceci n'est jamais appliqué en pratique car il y a perte d'efficacité due à la possibilité de tirer le même individu plusieurs fois. A supposer que les tirages avec remise se fassent bien indépendamment les uns des autres il est clair que, dans ce cas, les v.a. $X_1, X_2, \cdots, X_n$ sont i.i.d..

Remarquons intuitivement que, si le *taux de sondage* n/N est faible (par exemple un échantillon de taille 1000 dans la population française des individus âgés de 15 ans et plus), le sondage sans remise rejoint le sondage avec remise. Ceci justifie qu'en pratique on utilise les résultats de la théorie statistique classique développés dans les chapitres à venir, dans les situations de sondage. Disons que si n/N reste inférieur à 0,1 on a des approximations correctes, d'autant plus que d'autres approximations du même ordre de grandeur sont souvent inévitables dans la théorie des sondages elle-même.

Note 3.4 Nous avons considéré des tirages sans remises successifs. Il est équivalent, en pratique, de tirer simultanément n individus parmi les N individus de la population, chacun des $\binom{N}{n}$ échantillons possibles de taille n devant avoir la même probabilité $1/\binom{N}{n}$ d'être sélectionné.

3.8 Notation matricielle des vecteurs aléatoires

Pour établir les propriétés d'un p-uplet $(X_1, X_2, \cdots, X_p)$ de v.a. il est commode d'adopter la notation matricielle. Ainsi, on note :

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

le vecteur aléatoire de dimension p, à valeurs dans $\mathbb{R}^p$. On définit alors l'espérance mathématique de $\mathbf{X}$, notée $\mathbb{E}(\mathbf{X})$, par le vecteur des espérances mathématiques (si elles existent) :

$$\mathbb{E}(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix}.$$

Si les covariances des composantes prises 2 à 2 existent, la matrice d'élément (i, j) égal à $\mathrm{cov}(X_i, X_j)$ est appelée **matrice des variances-covariances** de $\mathbf{X}$ et nous la noterons $\mathbb{V}(\mathbf{X})$. Notons que cette matrice est symétrique et que ses éléments diagonaux sont les variances des composantes.

Soient maintenant $\mathbf{A}$ une matrice $(q \times p)$ et $\mathbf{c}$ un vecteur $(q \times 1)$. Alors la relation $\mathbf{Y} = \mathbf{AX} + \mathbf{c}$ définit un vecteur aléatoire $\mathbf{Y}$ à valeurs dans $\mathbb{R}^q$.

Proposition 3.13 *Soit* $\mathbf{X}$ *un vecteur aléatoire d'espérance* $\mathbb{E}(\mathbf{X})$ *et de matrice de variance-covariance* $\mathbb{V}(X)$ *et soit le vecteur aléatoire* $\mathbf{Y}$ *tel que* $\mathbf{Y} = \mathbf{AX} + \mathbf{c}$. *On a alors :*

$$\mathbb{E}(\mathbf{Y}) = \mathbf{A}\,\mathbb{E}(\mathbf{X}) + \mathbf{c}\,,$$
$$\mathbb{V}(\mathbf{Y}) = \mathbf{A}\,\mathbb{V}(\mathbf{X})\,\mathbf{A}^t\,.$$

Le symbole $\mathbf{A}^t$ désigne la matrice transposée de $\mathbf{A}$. Nous omettrons la démonstration qui ne présente pas de difficulté et offre peu d'intérêt.

Indiquons le cas particulier où $\mathbf{A}$ est le vecteur ligne $(1 \times p)$,

$$\mathbf{A} = (1 \quad 1 \quad \cdots \quad 1)\,,$$

pour lequel $\mathbf{Y}$ est la somme des composantes de $\mathbf{X}$. On obtient alors la généralisation de la proposition 3.9 à p v.a. quelconques :

$$V\left(\sum_{i=1}^{p} X_i\right) = \sum_{i=1}^{p} V(X_i) + 2\sum_{i<j} cov(X_i, X_j)\,,$$

où $\sum_{i<j}$ est une sommation sur tous les couples (X_i, X_j) avec $i < j$.

3.9 Loi de Gauss multivariée

La loi de Gauss, ou loi normale, pour une v.a. à valeurs dans $\mathbb{R}$ est la célèbre courbe en cloche (graphe de sa densité). Elle est décrite en détail en section 4.2.4. Indiquons simplement ici qu'il s'agit en fait d'une famille de lois dépendant de deux paramètres qui sont la moyenne μ et la variance σ^2 de chaque loi, d'où la notation $\mathcal{N}(\mu, \sigma^2)$. Toute fonction linéaire $aX + b$ d'une

v.a. gaussienne X est une v.a. gaussienne dont la moyenne et la variance se calculent par les règles générales vues en section 2.3 : $E(aX+b) = aE(X)+b$ et $V(aX+b) = a^2V(X)$. Par une transformation linéaire *ad hoc* on peut se ramener à la loi[2] $\mathcal{N}(0\,;1)$ appelée loi de Gauss centrée-réduite (celle fournie dans les tables). Nous montrons des propriétés analogues pour un vecteur aléatoire gaussien.

Un vecteur aléatoire gaussien $\mathbf{X}$ de dimension p est parfaitement défini par son vecteur des espérances noté $\boldsymbol{\mu}$ et sa matrice des variances-covariances notée $\boldsymbol{\Sigma}$. Sa loi est notée $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La densité conjointe de ses composantes au point $(x_1, x_2, \cdots, x_p) \in \mathbb{R}^p$ est :

$$f_{\mathbf{X}}(x_1, x_2, \cdots, x_p) = \frac{1}{(2\pi)^{p/2}(\det \boldsymbol{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\ \boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right\}$$

où $\det \Sigma$ dénote le déterminant de la matrice $\boldsymbol{\Sigma}$ et $\mathbf{x}$ dénote le vecteur colonne de composantes $x_1, x_2, \cdots, x_p$.

Notons que la matrice $\boldsymbol{\Sigma}$ doit être inversible. Sinon le vecteur aléatoire ne serait pas réellement de dimension p au sens où il y aurait au moins une liaison linéaire exacte entre ses composantes et, par conséquent, ce vecteur prendrait ses valeurs dans un sous-espace de dimension inférieure à p.

Un cas particulier important est celui de la loi dont le vecteur des espérances mathématiques est le vecteur nul $\mathbf{0}$ et la matrice des variances-covariances est la matrice identité $\mathbf{I}_p$, soit la loi $\mathcal{N}_p(\mathbf{0}, \mathbf{I_p})$ que nous appellerons *loi de Gauss p-variée centrée-réduite* par analogie avec la loi usuelle centrée-réduite $\mathcal{N}(0\,;1)$ dans le cas où $p = 1$. Pour cette loi, l'expression de la densité devient :

$$\frac{1}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} x_i^2\right\} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\}$$

qui se sépare en un produit de densités respectives à chacune des composantes, qui sont les densités marginales de celles-ci. En se reportant à la section 4.2.4 on peut voir que toutes les composantes ont une loi marginale $\mathcal{N}(0;1)$. Ainsi ces composantes sont indépendantes. Il est clair que ceci est également vrai si la matrice $\boldsymbol{\Sigma}$ est diagonale, à ceci près que chaque composante a une loi marginale gaussienne dont la moyenne est la composante correspondante du vecteur $\boldsymbol{\mu}$ et la variance est la valeur sur la position correspondante de la diagonale de $\boldsymbol{\Sigma}$. On a donc la proposition suivante.

Proposition 3.14 *Un vecteur aléatoire gaussien a ses composantes indépendantes si et seulement si la matrice des variances-covariances est diagonale, i.e. si et seulement si les covariances des composantes prises deux à deux sont toutes nulles.*

[2]Bien que la notation générique soit $\mathcal{N}(\mu, \sigma^2)$, nous aurons l'habitude de remplacer la virgule par un point-virgule lorsque l'on aura des nombres explicites afin d'éviter la confusion avec la virgule décimale.

Alors que nous avions vu en section 3.4 qu'une covariance nulle n'impliquait pas l'indépendance pour un couple de v.a. de façon générale, dans le cas gaussien **il y a équivalence entre indépendance et covariance (ou corrélation) nulle**.

Nous admettrons le théorème suivant très utile pour caractériser les vecteurs gaussiens.

Théorème 3.1 ***(théorème de caractérisation)*** *Un vecteur aléatoire est gaussien si et seulement si toute combinaison linéaire de ses composantes est une variable aléatoire gaussienne.*

On déduit immédiatement de cette caractérisation essentielle que si $\mathbf{X}$ est de loi $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ alors $\mathbf{Y} = \mathbf{A}\mathbf{X}$ est également un vecteur aléatoire gaussien puisque toute combinaison linéaire des composantes de $\mathbf{Y}$ est une combinaison linéaire des composantes de $\mathbf{X}$ et est donc gaussienne. De plus, le fait d'ajouter un vecteur de constantes $\mathbf{c}$ à un vecteur gaussien ne fait que déplacer la moyenne, la densité restant gaussienne avec $\boldsymbol{\mu} + \mathbf{c}$ se substituant à $\boldsymbol{\mu}$ dans l'expression générale donnée plus haut. En reprenant les résultats de la proposition 3.13, on en déduit la proposition suivante.

Proposition 3.15 *Soit* $\mathbf{X}$ *un vecteur aléatoire de loi* $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *et soit le vecteur aléatoire* $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{c}$ *où* $\mathbf{A}$ *est une matrice* $(q \times p)$ *de rang* q *et* $\mathbf{c}$ *un vecteur* $(q \times 1)$. *Alors* $\mathbf{Y}$ *est de loi* $\mathcal{N}_q(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}^t)$.

La condition que la matrice $\mathbf{A}$ soit de rang maximal (avec q nécessairement inférieur ou égal à p) s'impose pour que la matrice des variances-covariances $\mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}^t$ de $\mathbf{Y}$ soit inversible, i.e. que $\mathbf{Y}$ soit réellement de dimension q et non pas à valeurs dans un sous-espace de dimension inférieure.

On montre que l'on peut toujours par un choix judicieux de la matrice $\mathbf{A}$ et du vecteur $\mathbf{c}$ se ramener à un vecteur $\mathbf{Y}$ de loi p-variée centrée-réduite. En effet on peut d'abord ramener la moyenne à un vecteur nul en passant au vecteur aléatoire $\mathbf{X} - \boldsymbol{\mu}$. Puis, du fait que la matrice $\boldsymbol{\Sigma}$ est inversible et symétrique, il existe $\mathbf{C}$, une matrice $(p \times p)$ de rang p, telle que $\mathbf{C}\mathbf{C}^t = \boldsymbol{\Sigma}$. Considérons alors le vecteur $\mathbf{Y} = \mathbf{C}^{-1}(\mathbf{X} - \boldsymbol{\mu})$. Sa moyenne est évidemment nulle puisque $\mathbb{E}(\mathbf{Y}) = \mathbf{C}^{-1}\,\mathbb{E}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{C}^{-1}\mathbf{0} = \mathbf{0}$, et sa variance est :

$$\begin{aligned}\mathbb{V}(\mathbf{Y}) &= \mathbf{C}^{-1}\mathbb{V}(\mathbf{X})\,(\mathbf{C}^{-1})^t = \mathbf{C}^{-1}\boldsymbol{\Sigma}\,(\mathbf{C}^{-1})^t \\ &= \mathbf{C}^{-1}\mathbf{C}\mathbf{C}^t\,(\mathbf{C}^{-1})^t = \mathbf{I}_p\,.\end{aligned}$$

Remarques

1. En plus d'être symétrique et de rang maximal, la matrice $\boldsymbol{\Sigma}$ doit être définie strictement positive. C'est-à-dire que pour tout vecteur $\mathbf{v} \in \mathbb{R}^p$ non nul on a $\mathbf{v}^t\,\boldsymbol{\Sigma}\,\mathbf{v} > 0$. En effet une combinaison linéaire des composantes de $\mathbf{X}$ est de la forme $\mathbf{v}^t\mathbf{X}$, sa variance est $\mathbf{v}^t\,\boldsymbol{\Sigma}\,\mathbf{v}$, laquelle doit rester positive.

2. Pour qu'un couple de v.a. forme un couple gaussien, il ne suffit pas que chaque v.a. soit gaussienne. En d'autres termes les lois marginales peuvent être gaussiennes sans que la loi conjointe soit gaussienne sur $\mathbb{R}^2$. En revanche, toutes les lois marginales des composantes d'un vecteur gaussien sont des gaussiennes en tant que combinaison linéaire particulière donnant un coefficient 1 à cette composante et 0 à toutes les autres.

Par ailleurs si les composantes sont indépendantes et gaussiennes alors la loi conjointe est gaussienne multivariée.

3.10 Exercices

Exercice 3.1 Le tableau suivant représente la loi du couple (X, Y) : X nombre d'enfants dans un ménage, Y nombre de téléviseurs du ménage (pour un ménage pris au hasard dans une population de ménages ayant 1 à 3 enfants et 1 à 3 téléviseurs).

X \ Y	1	2	3
1	0,22	0,11	0,02
2	0,20	0,15	0,10
3	0,06	0,07	0,07

Calculer le coefficient de corrélation entre X et Y.

Exercice 3.2 Montrer que la covariance entre la somme et la différence de deux v.a. indépendantes et de même loi est toujours nulle.

Exercice 3.3 Soient X et Y deux v.a. indépendantes suivant une même loi de Bernoulli de paramètre p (voir section 4.1.2). Donner la loi de $X+Y$. Calculer $P(X+Y=0)$, $P(X-Y=0)$ et $P(X+Y=0,\ X-Y=0)$. Les deux v.a. $X+Y$ et $X-Y$ sont-elles indépendantes ? Que vaut leur covariance en application de l'exercice 3.2 ? Quelle conclusion générale en tirez-vous ?

Exercice 3.4 Soient X et Y deux v.a. indépendantes et soit $Z = X + Y$. Calculer $P(Z \leq z|X = x)$ et en déduire que $f_{Z|X=x}(z) = f_Y(z - x)$.

Déterminer la densité conjointe de Z et de X, et en déduire que $f_Z(z) = \int_{-\infty}^{+\infty} f_Y(z - x) f_X(x) dx$.

Donner la loi de $T = X - Y$.

Exercice 3.5 Déterminer la loi de la somme de deux v.a. indépendantes, continues uniformes sur $[0, 1]$.

Aide : on déterminera la zone du plan en coordonnées (x, y) définie par $\{(x, y) \,|\, x + y < z,\ 0 < x < 1,\ 0 < y < 1\}$ et, à partir de la loi conjointe, on calculera géométriquement, selon les différents cas pour z, $P(X + Y < z)$.

Exercice 3.6 On mesure la longueur et la largeur d'un terrain rectangulaire. La mesure de la longueur est une v.a. X de moyenne μ_X et de variance σ_X^2. La mesure de la largeur est une v.a. Y de moyenne μ_Y et de variance σ_Y^2. On suppose que ces deux mesures sont indépendantes. Quelle est l'espérance mathématique et la variance pour la mesure de la surface du terrain ?

Exercice 3.7 (marche aléatoire) On se situe sur un axe en une position initiale. On se déplace alors par étapes successives et indépendantes les unes des autres de la façon suivante. A chaque étape on fait un pas d'un mètre à droite (+1) avec probabilité p ou à gauche (-1) avec probabilité $1-p$. Soit X le déplacement à une étape quelconque. Calculer $E(X)$ et $V(X)$.

Soit Y l'éloignement de la position initiale après n étapes. Calculer $E(Y)$ et $V(Y)$.

Exercice 3.8 Soit un couple (X, Y) gaussien bivarié. On suppose que X et Y sont de moyenne nulle et on note σ_X^2 la variance de X, σ_Y^2 la variance de Y et ρ le coefficient de corrélation linéaire du couple. Établir que la densité conjointe du couple est :

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{x^2}{\sigma_X^2} - 2\rho\frac{xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2}\right]\right\}$$

(on notera que les courbes de niveau du graphe de $f_{X,Y}$ sont des ellipses de centre (0,0)).

Chapitre 4

Les lois de probabilités usuelles

Nous abordons ici une sorte de catalogue des lois les plus utilisées dans la modélisation statistique. Nous nous efforcerons de justifier l'utilité de ces lois en précisant le type de situations où elles sont appropriées. De façon générique et sauf mention expresse on notera X une v.a. qui suit la loi décrite. Chaque loi fera l'objet d'un symbole spécifique. Par exemple, la loi binomiale de paramètres n et p sera notée $\mathcal{B}(n,p)$ et on écrira $X \rightsquigarrow B(n,p)$ pour signifier que X suit cette loi.

4.1 Les lois discrètes

4.1.1 La loi uniforme discrète

L'ensemble des valeurs possibles est $\{1,2,3,\cdots,r\}$, r étant un paramètre de la loi. Uniforme signifie que chaque valeur reçoit la même probabilité $1/r$. En fait cette loi est peu utilisée en tant que modèle statistique, mais mérite d'être présentée en raison de sa simplicité. On la rencontre dans les jeux de hasard, par exemple dans le lancement d'un dé : X est le nombre de points obtenus et r est égal à 6. Si le dé est parfaitement symétrique chaque face, et donc chaque nombre de points, a la probabilité 1/6 d'apparaître.

Pour une v.a. X qui suit cette loi, on a :

$$E(X) = \frac{r+1}{2}$$
$$V(X) = \frac{1}{12}(r^2-1)\ .$$

En effet :

$$E(X) = \frac{1}{r}(1+2+\cdots+r) = \frac{1}{r}\frac{r(r+1)}{2} = \frac{r+1}{2}$$

et, sachant que $1^2+2^2+\cdots+r^2 = \frac{1}{6}r(r+1)(2r+1)$, on peut calculer la variance par la formule de décentrage :

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= \frac{1}{r}(1^2 + 2^2 + \cdots + r^2) - \left(\frac{r+1}{2}\right)^2 \\ &= \frac{1}{6}(r+1)(2r+1) - \left(\frac{r+1}{2}\right)^2 = \frac{1}{12}(r^2-1)\ . \end{aligned}$$

Ainsi, pour le jet d'un dé, l'espérance mathématique du nombre de points est $\frac{7}{2}$ et sa variance $\frac{35}{12}$.

4.1.2 Loi de Bernoulli $\mathcal{B}(p)$

C'est **la loi la plus simple** que l'on puisse envisager puisqu'il n'y a que deux valeurs possibles, codées 1 et 0. On note p la probabilité associée à la valeur 1, p étant le paramètre de la loi (la probabilité $1-p$ associée à la valeur 0 est souvent notée q dans les ouvrages, mais nous jugeons cela superflu). On écrit $X \rightsquigarrow B(p)$ et donc :

$$X \rightsquigarrow B(p) \iff X \left\{ \begin{array}{lcc} \text{valeurs possibles} & 1 & 0 \\ \text{probabilités} & p & 1-p\ . \end{array} \right.$$

On peut écrire la fonction de probabilité de la façon suivante :

$$P(X=x) = p^x(1-p)^{1-x} \quad , \quad x \in \{0,1\}.$$

On a $E(X) = p$ et $V(X) = p(1-p)$ puisque :

$$\begin{aligned} &E(X) = 0 \times (1-p) + 1 \times p = p \\ &V(X) = E(X^2) - (E(X))^2 = 0^2 \times (1-p) + 1^2 \times p - p^2 = p(1-p)\,. \end{aligned}$$

La fonction génératrice des moments est :

$$\begin{aligned} \Psi_X(t) &= E(e^{tX}) = e^{t.1}p + e^{t.0}(1-p) \\ &= pe^t + (1-p)\,. \end{aligned}$$

En pratique la v.a. X sera utilisée comme **fonction indicatrice d'un événement** donné au cours d'une expérience aléatoire (par exemple avoir un

appareil tombant en panne avant l'expiration de la garantie, être infecté au cours d'une épidémie, être bénéficiaire pour une entreprise). X prend la valeur 1 si l'événement se produit et 0 s'il ne se produit pas à l'issue de l'expérience. Dans ce contexte, p **représente la probabilité de l'événement considéré.** La v.a. X sera une variable de comptage lors de répétitions de l'expérience constituant le processus de Bernoulli décrit ci-après et conduisant notamment à la loi binomiale.

Par convention la réalisation de l'événement sera appelée «*succès*» et sera codée 1, sa non-réalisation sera appelée «*échec*» et sera codée 0.

4.1.3 Le processus de Bernoulli et la loi binomiale $\mathcal{B}(n, p)$

Le processus consiste en une suite de répétitions de l'expérience aléatoire de Bernoulli, toutes ces répétitions successives étant indépendantes les unes des autres. La probabilité de succès à chaque répétition est p.

Un processus de Bernoulli est donc modélisé par une suite $X_1, X_2, X_3, \ldots$ de v.a. i.i.d., chacune de loi $\mathcal{B}(p)$. Dans ce processus on peut s'intéresser à différents types de comptages, menant à différentes lois. Nous verrons les plus courants : comptage des succès en s'arrêtant à un nombre de répétitions fixé à l'avance (loi binomiale), comptage des échecs avant d'atteindre le premier succès (loi géométrique) ou le r-ième succès (loi binomiale négative).

La *loi binomiale* est la loi de la v.a. X correspondant au **nombre de succès au cours de n répétitions** du processus. Elle est omniprésente en statistique. L'application la plus fréquente se situe dans le domaine des sondages. Ayant sélectionné au hasard n individus dans une grande population (voir le sondage aléatoire simple en section 3.7) on peut «estimer» la **proportion p d'individus ayant un *caractère*[1] donné** (succès). Si le taux de sondage est faible, on a vu que l'on pouvait admettre que le tirage sans remise est très proche du tirage avec remise. Pour ce dernier la probabilité de succès à chaque tirage est p et il y a indépendance des tirages. La v.a. X correspond au nombre d'individus ayant le caractère d'intérêt parmi n individus sélectionnés.

La loi binomiale a deux paramètres n et p, et l'ensemble des valeurs possibles est $\{0, 1, 2, \cdots, n\}$. Calculons directement la probabilité $p(x)$ d'obtenir x succès parmi n répétitions.

Toute suite contenant x succès et $n-x$ échecs a une probabilité $p^x(1-p)^{n-x}$ en raison de l'indépendance des répétitions successives, et ceci quel que soit l'ordre d'apparition des succès et des échecs. Imaginons que nous écrivions la succession des résultats avec une séquence de lettres S et E (succès, échec). Combien y-a-t-il d'écritures possibles ? Une suite particulière étant parfaitement définie par les positions occupées par les x lettres S, il suffit de dénombrer

[1] Le mot «caractère» est à prendre dans un sens élargi. Ce peut être, par exemple, l'acquiescement à une opinion proposée dans un questionnaire.

combien il y a de choix de x positions parmi n positions. C'est le nombre de combinaisons à x éléments que l'on peut former à partir de n éléments :

$$\binom{n}{x} = \frac{n!}{p!(n-p)!}\,.$$

D'où $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$, toutes ces suites étant distinctes et donc incompatibles.

Définition 4.1 *On dit que la v.a. discrète X suit une loi binomiale $\mathcal{B}(n,p)$ si sa fonction de probabilité est :*

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}\ ,\ x = 0, 1, 2, \cdots, n\,.$$

Proposition 4.1 *Soit $X_1, X_2, \cdots, X_n$ une suite de v.a. i.i.d. de loi $\mathcal{B}(p)$, alors $S_n = \sum_{i=1}^{n} X_i$ suit une loi $\mathcal{B}(n,p)$.*

Ceci est la traduction du nombre de comptage des succès à travers la variable indicatrice de Bernoulli.

De cette proposition on déduit que la somme de deux v.a. indépendantes de lois respectives $\mathcal{B}(n_1,p)$ et $\mathcal{B}(n_2,p)$ est une v.a. de loi $\mathcal{B}(n_1+n_2,p)$. En effet cette somme peut être considérée comme celle de n_1+n_2 répétitions indépendantes du processus de Bernoulli avec probabilité p de succès.

Proposition 4.2 *Soit $X \rightsquigarrow B(n,p)$, alors :*

$$\begin{aligned} E(X) &= np \\ V(X) &= np(1-p) \\ \Psi_X(t) &= \left[pe^t + (1-p)\right]^n\,. \end{aligned}$$

Démonstration : comme X peut être vue comme la somme de n v.a. indépendantes $X_1, X_2, \cdots, X_n$ de même loi $\mathcal{B}(p)$, il suffit d'appliquer la proposition fondamentale 3.11 sur la somme de n v.a. i.i.d., avec $\mu = p$ et $\sigma^2 = p(1-p)$ qui sont respectivement la moyenne et la variance de la loi $\mathcal{B}(p)$, pour obtenir la moyenne et la variance de X . En appliquant le résultat de la proposition 3.12 on obtient sa fonction génératrice des moments. On peut vérifier, à titre d'exercice, que $\Psi_X^{'}(0) = np$ (voir section 2.5). □

4.1.4 Les lois géométrique $\mathcal{G}(p)$ et binomiale négative $\mathcal{BN}(r,p)$

Soit un processus de Bernoulli de paramètre p. La *loi géométrique* $\mathcal{G}(p)$, ou loi de Pascal, est la loi de la v.a. X «nombre d'échecs avant de parvenir au premier succès». L'ensemble des valeurs possibles est $\mathbb{N}$ et la fonction de probabilité est :

$$p(x) = p(1-p)^x,\ x \in \mathbb{N}\ ,$$

car il n'y a qu'une séquence possible : x échecs suivis d'un succès.
On a alors :

$$E(X) = \frac{1-p}{p}$$
$$V(X) = \frac{1-p}{p^2}$$
$$E(e^{tX}) = \frac{p}{1-(1-p)e^t}\,.$$

Démonstration : la fonction génératrice s'écrit :

$$E(e^{tX}) = \sum_{k=0}^{\infty} e^{tk} p(1-p)^k = p\sum_{k=0}^{\infty}[(1-p)e^t]^k = \frac{p}{1-(1-p)e^t}\ .$$

Elle est définie si $(1-p)e^t < 1$ ou $t < -\ln(1-p)$, donc au voisinage de 0. La dérivée première de cette expression est :

$$\frac{p(1-p)e^t}{[1-(1-p)e^t]^2}$$

et sa valeur pour $t=0$ est $(1-p)/p$ qui correspond à la moyenne. En prenant la dérivée seconde au point 0, on obtient $E(X^2) = (1-p)(2-p)/p^2$, puis $V(X)$ par la formule de décentrage. □

La *loi binomiale négative* est une généralisation de la loi géométrique où l'on considère X «nombre d'échecs avant de parvenir au r-ième succès». Sa fonction de probabilité est :

$$p(x) = \binom{r+x-1}{x} p^r (1-p)^x,\ x \in \mathbb{N}.$$

En effet pour toute séquence de x échecs et r succès la probabilité est $p^r(1-p)^x$. Sachant que le dernier résultat de la séquence doit être un succès, il reste à dénombrer les séquences avec x échecs et $r-1$ succès ce qui revient à dénombrer les possibilités de choix de x positions parmi $x+r-1$ positions, soit $\binom{r+x-1}{x}$.

On a alors :

$$E(X) = \frac{r(1-p)}{p}$$
$$V(X) = \frac{r(1-p)}{p^2}$$
$$\Psi_X(t) = [\frac{p}{1-(1-p)e^t}]^r \text{ (au voisinage de 0) .}$$

Ceci peut être établi grâce aux propositions 3.11 et 3.12, en remarquant qu'une v.a. $\mathcal{BN}(r,p)$ peut être considérée comme une somme de r v.a. indépendantes de loi $\mathcal{G}(p)$. En effet toute séquence à r succès, dont un succès final, peut être vue comme une suite de r séquences du type de celles de la loi géométrique.

Certains auteurs préfèrent à X la v.a. Y «nombre **total** de répétitions pour atteindre le r-ième succès». On a donc $Y = X + r$, avec :

$$p_Y(x) = \binom{x-1}{x-r} p^r (1-p)^{x-r},\ x = r, r+1, ...$$

et :
$$E(Y) = \frac{r}{p},\ \ V(Y) = \frac{r(1-p)}{p^2}.$$

4.1.5 La loi hypergéométrique $\mathcal{H}(N, M, n)$

Soit un ensemble de N individus dont M possèdent un certain caractère, que nous appellerons «succès» par analogie avec la loi binomiale, et $N - M$ ne la possèdent pas. On effectue un **tirage aléatoire sans remise** de n individus dans cet ensemble. On entend par là que chacun des $\binom{N}{n}$ échantillons de taille n possibles a la même probabilité $1/\binom{N}{n}$d'être sélectionné (voir note 3.4). On considère la v.a. X «nombre de succès observés parmi les n individus». On a alors la fonction de probabilité :

$$p(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},\ \ x = 0, 1, ..., n.$$

Le numérateur correspond au nombre de choix de x individus parmi M et $n-x$ parmi $N - M$. Avec comme valeurs possibles $\{0, 1, ..., n\}$ nous supposons que M et $N-M$ sont supérieurs à n. Toutefois si $n > M$ alors la plus grande valeur possible est M et si $n > N - M$ la plus petite valeur possible est $n - (N - M)$. Nous pouvons garder la formule générale ci-dessus en convenant que $\binom{a}{b} = 0$ si $a < b$.

On démontre (le résultat étant intuitif pour la moyenne) que :

$$E(X) = n\frac{M}{N}$$
$$V(X) = n\frac{M}{N}(1 - \frac{M}{N})\frac{N-n}{N-1} .$$

On peut faire le rapprochement avec le processus de Bernoulli et la loi binomiale qui correspondent au tirage aléatoire avec remise. Le nombre de succès suit une loi $\mathcal{B}(n,p)$ avec $p = M/N$. Dans les deux situations la moyenne est identique alors que, pour la loi hypergéométrique, la variance reçoit un «*facteur correctif de sans remise*» égal à $(N-n)/(N-1)$. Clairement les deux situations se rapprochent si le «taux de sondage» n/N diminue. Plus formellement on peut montrer que, pour tout x, $p(x)$ tend vers l'expression correspondante $\binom{n}{x}p^x(1-p)^{n-x}$ de la loi binomiale quand $N \to \infty$ et $(M/N) \to p$.

4.1.6 La loi multinomiale

Il s'agit d'une extension de la situation binaire (succès, échec) de la loi binomiale, à une situation «multinaire» à c catégories, de probabilités respectives $p_1, p_2, ..., p_c$ avec $\sum_{k=1}^{c} p_k = 1$.

On s'intéresse aux fréquences observées $N_1, N_2, ..., N_c$ des différentes catégories au cours de n observations répétées indépendantes. La fonction de probabilité conjointe des v.a. $N_1, N_2, ..., N_c$ est :

$$P(N_1 = n_1, N_2 = n_2, ..., N_c = n_c) = \frac{n!}{n_1!n_2!...n_c!} p_1^{n_1} p_2^{n_2} ... p_c^{n_c}$$

si tous les n_k (k de 1 à c) appartiennent à $\{0, 1, 2, ..., n\}$ et vérifient la contrainte $\sum_{k=1}^{c} n_k = n$, la probabilité étant nulle sinon. Ceci s'établit par le même type de raisonnement que pour la loi binomiale. Le terme $p_1^{n_1} p_2^{n_2} \cdots p_c^{n_c}$ correspond à la probabilité de toute série de n répétitions avec n_1 d'entre elles donnant la catégorie 1, n_2 donnant la catégorie 2,... et n_c donnant la catégorie c. Le terme avec les factoriels correspond au nombre de séries de ce type possibles (nombre de façons d'occuper n_1 positions parmi les n successions pour la catégorie 1, n_2 positions pour la catégorie 2, $\cdots$, n_c pour la catégorie c).

La loi marginale de N_k est clairement la loi $B(n, p_k)$. Par ailleurs on démontre que $cov(N_k, N_l) = -np_kp_l$. On peut donc écrire l'espérance et la matrice des variances-covariances du vecteur aléatoire $\mathbf{N} = (N_1, N_2, ..., N_c)$ à valeurs dans $\{0, 1, 2, ..., n\}^c$ avec la contrainte $\sum_{k=1}^{c} N_k = n$:

$$\mathbb{E}(\mathbf{N}) = \begin{pmatrix} np_1 \\ np_2 \\ \vdots \\ np_c \end{pmatrix}, \ \mathbb{V}(\mathbf{Y}) = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \cdots & -np_1p_c \\ -np_1p_2 & np_2(1-p_2) & \cdots & -np_2p_c \\ \vdots & \vdots & \cdots & \vdots \\ -np_1p_c & -np_2p_c & \cdots & np_c(1-p_c) \end{pmatrix}.$$

Cette loi est à la base de l'étude des *variables catégorielles* du chapitre 10.

4.1.7 Le processus et la loi de Poisson $\mathcal{P}(\lambda)$

On considère un processus d'occurrences d'un événement donné sur l'échelle du temps, par exemple l'arrivée des appels à un standard téléphonique. Pour un

temps $t > 0$ fixé (à partir d'une certaine origine des temps) on définit la variable aléatoire $X(t)$ «nombre d'occurrences dans l'intervalle $]0, t]$». Par commodité on pose :

$$p_k(t) = P(X(t) = k), \text{ où } k \in \mathbb{N}.$$

En bref, on dit qu'on a un *processus de Poisson* si :

- il y a une invariance temporelle, à savoir que $p_k(t)$ ne dépend pas de l'origine des temps, mais dépend uniquement de la longueur t de l'intervalle, quels que soient k et t ;
- il y a indépendance des nombres d'occurrences pour deux intervalles disjoints ;
- pour un très petit intervalle la probabilité d'avoir deux occurrences ou plus est négligeable devant la probabilité d'avoir une occurrence exactement et cette dernière est proportionnelle à la longueur de cet intervalle. Plus formellement :

$$p_1(h) = \lambda h + o(h)$$
$$\sum_{k=2}^{\infty} p_k(h) = o(h)$$

où, rappelons-le, $o(h)$ est une fonction telle que $\dfrac{o(h)}{h} \to 0$ quand $h \to 0$. Le paramètre $\lambda > 0$ caractérise l'intensité de fréquence des occurrences.

Sous ces hypothèses on démontre que :

$$p_k(t) = \frac{e^{-\lambda t}(\lambda t)^k}{k!},\ k \in \mathbb{N}.$$

La *loi de Poisson* est la loi du **nombre d'occurrences dans une unité de temps**, donc pour $t = 1$ dans les formulations ci-dessus. Par conséquent on dit que la v.a. X suit une loi de Poisson $\mathcal{P}(\lambda)$ si sa fonction de probabilité est :

$$p(k) = \frac{e^{-\lambda}\lambda^k}{k!},\ k \in \mathbb{N}.$$

Sachant que $\displaystyle\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$, la somme des probabilités est bien égale à 1. On a alors :

$$\begin{aligned} E(X) &= \lambda \\ V(X) &= \lambda \\ \Psi_X(t) &= e^{\lambda(e^t - 1)}. \end{aligned}$$

Le paramètre λ est donc le **nombre moyen d'occurrences par unité de temps** pour le processus. Les démonstrations sont simples :

$$E(X) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \lambda\frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda\sum_{j=0}^{\infty}\frac{e^{-\lambda}\lambda^j}{j!} = \lambda$$

$$\begin{aligned} E(X^2) &= \sum_{k=0}^{\infty} k^2\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=2}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} + \sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} \\ &= \lambda^2\sum_{k=2}^{\infty}\frac{e^{-\lambda}\lambda^{k-2}}{(k-2)!} + \lambda\sum_{k=1}^{\infty}\frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda^2 + \lambda \end{aligned}$$

d' où $V(X) = \lambda$,

$$\Psi_X(t) = \sum_{k=0}^{\infty}\frac{e^{tk}e^{-\lambda}\lambda^k}{k!} = \sum_{k=0}^{\infty}\frac{e^{-\lambda}(\lambda e^t)^k}{k!} = e^{\lambda(e^t-1)}\sum_{k=0}^{\infty}\frac{e^{-\lambda e^t}(\lambda e^t)^k}{k!} = e^{\lambda(e^t-1)}.$$

Remarques

- On verra plus loin la loi exponentielle qui est celle du temps s'écoulant entre deux occurrences successives.
- On a la propriété additive suivante : soient $X_1 \rightsquigarrow \mathcal{P}(\lambda_1)$ et $X_2 \rightsquigarrow \mathcal{P}(\lambda_2)$ indépendante de X_1, alors $X_1+X_2 \rightsquigarrow \mathcal{P}(\lambda_1+\lambda_2)$. Ceci se voit directement en appliquant la proposition 3.10 sur la fonction génératrice d'une somme.

Approximation de la loi binomiale par la loi de Poisson

Si l'on choisit une unité de temps suffisamment petite pour que la probabilité d'avoir plus d'une occurrence devienne négligeable on voit que le processus de Poisson peut être rapproché d'un processus de Bernoulli par discrétisation de l'écoulement continu du temps en unités successives.

Montrons que la fonction de probabilité de la loi $\mathcal{P}(\lambda)$ est équivalente à celle de la loi $\mathcal{B}(n,p)$ quand $n \to \infty$ et $p \to 0$ de façon que $np \to \lambda$. On a vu (section 4.1.3) que la fonction génératrice des moments de la loi $\mathcal{B}(n,p)$ est $\Psi(t) = [pe^t + (1-p)]^n$. D'où :

$$\begin{aligned} \ln\Psi(t) &= n\ln\left[pe^t + (1-p)\right] \\ &= n\ln\left[1 + p(e^t-1)\right] \\ &= n[p(e^t-1) + o(p)]\,. \end{aligned}$$

Comme np tend vers λ, $\ln\Psi(t)$ tend vers $\lambda(e^t-1)$. Par suite $\Psi(t)$ tend vers $\exp\{\lambda(e^t-1)\}$ qui est la fonction génératrice de la loi $\mathcal{P}(\lambda)$.

Ceci a un intérêt pratique pour approcher la loi binomiale lorsque l'événement «succès» est rare (p est petit), avec un grand nombre de répétitions.

On considère que, si $n \geq 50$ et $p \leq 0,1$, la loi binomiale $\mathcal{B}(n,p)$ est approchée de façon tout à fait satisfaisante par la loi de Poisson de paramètre $\lambda = np$. On peut aussi utiliser une telle approximation si l'événement «succès» est très fréquent ($p \geq 0,9$) en intervertissant succès et échec.

Exemple 4.1 La probabilité pour qu'un réacteur d'avion d'un certain type connaisse une panne avant sa première révision est 1/1000. Sachant qu'une compagnie d'aviation possède sur ses avions 100 réacteurs de ce type calculons la probabilité qu'elle ne rencontre pas plus de deux problèmes avec ces réacteurs avant la première révision. Le nombre de réacteurs à problème est une v.a. X de loi $\mathcal{B}(100\,;0{,}001)$ qui peut être approchée par la loi $\mathcal{P}(0{,}10)$. Donc :

$$P(X \leq 2) \simeq e^{-0,1} + \frac{e^{-0,1}.0{,}1}{1!} + \frac{e^{-0,1}.(0{,}1)^2}{2!} = 0{,}99985.$$

■

Le modèle de Poisson s'applique dans de nombreuses **situations de comptages** par unité de temps ou par unité de surface : nombre de sinistres par an pour un assuré, problèmes de files d'attente (arrivées à un guichet, nombre de personnes servies), particules émises par une source radioactive. Pour un comptage par unité de surface (par exemple le nombre de couples d'une espèce d'oiseaux nichant par quadrat d'une forêt), le modèle correspond à une répartition spatiale au hasard.

4.2 Les lois continues

4.2.1 La loi continue uniforme $\mathcal{U}[a,b]$

On dit que X suit une loi uniforme sur l'intervalle fini $[a,b]$ si **sa densité est constante** sur $[a,b]$ et nulle à l'extérieur de cet intervalle, soit :

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon} \end{cases} .$$

Nous en avons déjà vu une illustration en section 1.4.
Sa fonction de répartition est :

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \dfrac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } x > b \end{cases} .$$

On peut aisément vérifier que :

$$E(X) = \frac{a+b}{2} \ ,$$
$$V(X) = \frac{(b-a)^2}{12} \ .$$

La loi uniforme de référence est la loi $\mathcal{U}[0\,,1]$ correspondant aux *générateurs de nombres au hasard* des logiciels (fonction «RANDOM» ou «ALEA»). A partir d'un tel générateur on peut produire des nombres au hasard sur $[a, b]$ par la transformation $y = (b-a)x + a$. Nous verrons en section 4.3 comment simuler une loi quelconque à partir de ces «nombres au hasard».

4.2.2 La loi exponentielle $\mathcal{E}(\lambda)$

Comme mentionné dans les remarques sur le processus de Poisson, la loi exponentielle correspond à la variable aléatoire X du temps s'écoulant entre deux occurrences successives lors d'un tel processus. Avec les notations de la section 4.1.7 la probabilité qu'il n'y ait aucune occurrence dans un intervalle de temps de longueur t est égale à $p_0(t) = e^{-\lambda t}$, d'où $P(X > t) = e^{-\lambda t}$ et l'expression de la fonction de répartition $P(X \leq t)$:

$$F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{si } t \geq 0 \\ 0 & \text{si } t < 0 \end{cases} \ ,$$

puis de la densité, par dérivation :

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

On a déjà montré (voir section 2.5) que :

$$E(X) = \frac{1}{\lambda}$$
$$V(X) = \frac{1}{\lambda^2}$$
$$\Psi_X(t) = \frac{\lambda}{\lambda - t} \ .$$

Logiquement, puisque λ est le nombre moyen d'occurrences par unité de temps, $\frac{1}{\lambda}$ est la durée moyenne entre deux occurrences successives. On reparamétrise souvent la loi en posant $\theta = 1/\lambda$, d'où :

$$f(x) = \frac{1}{\theta}\, e^{-\dfrac{x}{\theta}}, \quad x \geq 0,$$

qui met en évidence sa moyenne θ, la variance étant alors θ^2.

La loi exponentielle est également le modèle de **durée de vie pour un système idéal sans usure**, $\frac{1}{\lambda}$ étant l'espérance de vie du système. En effet on peut voir que l'âge du système ne joue aucun rôle quant aux chances de survie à un horizon donné puisque :

$$\begin{aligned} P(X > t+h | X > t) &= \frac{P((X > t+h) \cap (X > t))}{P(X > t)} \\ &= \frac{P(X > t+h)}{P(X > t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} \ , \end{aligned}$$

qui ne dépend pas de t.

4.2.3 La loi gamma $\Gamma(r, \lambda)$

Soit $X_1, X_2, ..., X_r$ une suite de r variables aléatoires i.i.d. de loi $\mathcal{E}(\lambda)$ et soit $T = \sum_{i=1}^{r} X_i$. On démontre (voir exercices) que T suit une loi de densité :

$$f(x) = \begin{cases} \dfrac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \ ,$$

laquelle définit la loi $\Gamma(r, \lambda)$. Des propriétés des sommes de v.a. i.i.d. on déduit immédiatement :

$$\begin{aligned} E(T) &= \frac{r}{\lambda} \\ V(T) &= \frac{r}{\lambda^2} \\ \Psi_T(t) &= (\frac{\lambda}{\lambda - t})^r, \quad \text{si } t < \lambda \, . \end{aligned}$$

Vérifions que la densité ci-dessus est bien celle qui conduit à cette fonction génératrice des moments :

$$\begin{aligned} \Psi_T(t) = E(e^{tT}) &= \int_0^{+\infty} e^{tx} \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x} dx \\ &= \frac{\lambda^r}{(r-1)!} \int_0^{+\infty} x^{r-1} e^{-(\lambda - t)x} dx \\ &= \frac{\lambda^r}{(r-1)!} \frac{1}{(\lambda - t)^r} \int_0^{+\infty} u^{r-1} e^{-u} du = \frac{\lambda^r}{(\lambda - t)^r} \end{aligned}$$

en vertu de la relation classique $\int_0^{+\infty} u^{r-1} e^{-u} du = (r-1)!$.

La loi $\Gamma(r, \lambda)$ modélise en particulier le temps séparant une occurrence de la r-ième suivante dans un processus de Poisson. Elle joue un rôle similaire à celui de la loi binomiale négative dans le processus de Bernoulli.

On peut généraliser la loi $\Gamma(r, \lambda)$ à une valeur de r non entière (mais strictement positive) en remplaçant, dans la densité, l'expression $(r-1)!$ par la fonction gamma d'Euler définie, pour tout réel positif, par $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$, dont la loi a hérité son nom.

La fonction de répartition de cette loi n'est pas explicite et nécessite le recours à un logiciel (ou des tables, mais celles-ci ne sont pas très courantes).

4.2.4 La loi de Gauss ou loi normale $\mathcal{N}(\mu, \sigma^2)$

Il s'agit, comme on sait, de la loi de probabilité fondamentale de la statistique en raison du théorème central limite que nous verrons en section 5.8.3.

On dit que la variable aléatoire X suit *une loi de Gauss*, ou *loi normale*, notée $\mathcal{N}(\mu, \sigma^2)$, si elle a pour densité :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}, \quad x \in \mathbb{R}.$$

Les paramètres sont notés μ et σ^2 du fait qu'ils correspondent respectivement à la moyenne et à la variance de la loi (voir la démonstration ci-après), σ étant donc son écart-type. Le graphe de la densité est la fameuse courbe en cloche symétrique autour de la valeur μ.

Pour $\mu = 0$ et $\sigma^2 = 1$ on a la *loi de Gauss centrée-réduite* $\mathcal{N}(0\,;1)$ dont la fonction de répartition, notée Φ, est donnée dans les tables statistiques usuelles :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{z^2}{2}} dz.$$

Montrons que si $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ alors sa transformée centrée-réduite $Z = \frac{X-\mu}{\sigma}$ suit la loi $\mathcal{N}(0\,;1)$. Soient F_X et F_Z les fonctions de répartition respectives de X et de Z, alors :

$$F_Z(Z \leq z) = P(Z \leq z) = P(\frac{X-\mu}{\sigma} \leq z) = P(X \leq \mu + z\sigma) = F_X(\mu + \sigma z),$$

et en dérivant F_Z et F_X par rapport à z, on a :

$$f_Z(z) = \sigma f_X(\mu + z\sigma) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(\mu + z\sigma - \mu)^2}{\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

En inversant le raisonnement on montre que si $Z \rightsquigarrow \mathcal{N}(0\,;1)$ alors $X = \mu + \sigma Z \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$ et, plus généralement, ceci implique que **toute fonction linéaire d'une v.a. gaussienne est une v.a. gaussienne**.

En vertu de cette propriété le calcul de $P(X \leq x)$ se ramène à un calcul de probabilité sur la variable gaussienne centrée-réduite. Mettons en évidence la règle de calcul par la proposition suivante.

Proposition 4.3 *Soit $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, alors :*

$$P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \quad \textit{où } Z \rightsquigarrow \mathcal{N}(0;1),$$

soit encore : $$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Exemple 4.2 Soit $X \rightsquigarrow \mathcal{N}(10\,;4)$. Calculons, par exemple, $P(X \leq 13)$:

$$P(X \leq 13) = P(Z \leq \frac{13-10}{2}) = P(Z \leq 1{,}5) = \Phi(1{,}5) = 0{,}9332$$

par lecture de la table de la loi normale centrée-réduite.

En lecture inverse déterminons le quantile d'ordre 0,95 de la loi de X. Pour Z on lit dans la table que le quantile d'ordre 0,95 est 1,645 , i.e. $P(Z \leq 1{,}645) = 0{,}95$. D'où $P(\frac{X-10}{2} \leq 1{,}645) = 0{,}95$ et $P(X \leq 10 + 1{,}645 \times 2) = 0{,}95$. Pour la loi de X le quantile est donc $10 + 1{,}645 \times 2 = 13{,}29$. ■

D'une façon générale $P(X \leq x)$ est obtenu en lisant la probabilité d'être inférieur à $\dfrac{x-\mu}{\sigma}$ dans la table et le quantile d'ordre α de la loi de X est égal à $\mu + z_\alpha\, \sigma$ où z_α est le quantile d'ordre α de la table, i.e. tel que $\Phi(z_\alpha) = \alpha$.

Calculons la fonction génératrice des moments de $Z \rightsquigarrow \mathcal{N}(0;1)$:

$$\begin{aligned}\Psi_Z(t) = E(e^{tZ}) &= \int_{-\infty}^{+\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(z-t)^2} e^{\frac{t^2}{2}} dz = \frac{e^{\frac{t^2}{2}}}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} du\,.\end{aligned}$$

D'où :

$$\Psi_Z(t) = e^{\frac{t^2}{2}}.$$

Par cette fonction, vérifions que la moyenne et la variance de Z sont bien, respectivement, 0 et 1 :

$$\begin{aligned}&\Psi'_Z(t) = te^{\frac{t^2}{2}}, \quad E(Z) = \Psi'_Z(0) = 0\,, \\ &\Psi''_Z(t) = (1+t^2)e^{\frac{t^2}{2}}, \quad V(Z) = E(Z^2) = \Psi''_Z(0) = 1.\end{aligned}$$

Pour $X = \mu + \sigma Z$ on a donc $E(X) = \mu$ et $V(X) = \sigma^2 V(Z) = \sigma^2$, ce qui justifie la notation $\mathcal{N}(\mu, \sigma^2)$.

On peut directement accéder aux moments de tous ordres par le développement en série entière de $e^{\frac{t^2}{2}}$ (voir la note 2.3) :

$$e^{\frac{t^2}{2}} = \sum_{s=0}^{\infty} \frac{(\frac{t^2}{2})^s}{s!} = \sum_{s=0}^{\infty} \frac{1}{2^s s!} t^{2s}$$

d'où :

$$\mu_{2s} = \frac{(2s)!}{2^s s!}.$$

Du fait que la densité est une fonction paire les moments d'ordres impairs sont nuls. On notera, pour mémoire, que $\mu_4 = 3$.

Pour $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$, comme $(X-\mu)^r = \sigma^r Z^r$, on obtient immédiatement les moments centrés :

$$\mu'_{2s} = \sigma^{2s} \frac{(2s)!}{2^s s!}$$

et, en particulier, $\mu'_4 = 3\sigma^4$.

De plus :

$$\Psi_X(t) = \exp(t\mu + \sigma^2 \frac{t^2}{2}),$$

car :

$$E(e^{tX}) = E(e^{t(\mu+\sigma Z)}) = e^{t\mu} E(e^{(t\sigma)Z}) = e^{t\mu} e^{\frac{(t\sigma)^2}{2}}.$$

Nous donnons maintenant une proposition essentielle pour les développements statistiques.

Proposition 4.4 *Toute combinaison linéaire de v.a. gaussiennes indépendantes est une variable aléatoire gaussienne.*

Démonstration : il suffit de démontrer cela avec deux v.a., l'extension à plusieurs v.a. se faisant de proche en proche. De plus, on a vu que si X est gaussienne alors $Y = aX$ est gaussienne. Il suffit donc de démontrer la proposition pour $Y_1 + Y_2$, où Y_1 et Y_2 sont indépendantes. Soient $Y_1 \rightsquigarrow \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_2 \rightsquigarrow \mathcal{N}(\mu_2, \sigma_2^2)$. Selon la proposition 3.10 on a :

$$\begin{aligned}\Psi_{Y_1+Y_2}(t) &= \Psi_{Y_1}(t)\Psi_{Y_2}(t:) = e^{t\mu_1 + \frac{\sigma_1^2}{2}t^2} e^{t\mu_2 + \frac{\sigma_2^2}{2}t^2} \\ &= e^{t(\mu_1+\mu_2) + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}\end{aligned}$$

qui est la fonction génératrice des moments de la loi $\mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. □

Notons que la proposition n'est pas vraie pour des v.a. dépendantes. Ainsi deux v.a. peuvent être marginalement gaussiennes sans pour autant que toute

combinaison linéaire de celles-ci soit gaussienne, car cela dépend de la nature de leur loi conjointe.

Quelques valeurs clés de la loi de Gauss

A partir de la lecture dans la table de la loi de Gauss centrée-réduite des quantiles d'ordres 0,975 ; 0,995 et 0,9995, soit :

$$\begin{aligned}\Phi(0{,}975) &= 1{,}96\\ \Phi(0{,}995) &= 2{,}57\\ \Phi(0{,}9995) &= 3{,}30\ ,\end{aligned}$$

on déduit ces intervalles de dispersion autour de la moyenne pour $X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$:

$$\begin{aligned}P(\mu - 1{,}96\,\sigma < X < \mu + 1{,}96\,\sigma) &= 0{,}95\\ P(\mu - 2{,}57\,\sigma < X < \mu + 2{,}57\,\sigma) &= 0{,}99\\ P(\mu - 3{,}30\,\sigma < X < \mu + 3{,}30\,\sigma) &= 0{,}999\,.\end{aligned}$$

La première égalité est souvent résumée en disant que la probabilité d'obtenir une valeur dans l'intervalle «moyenne plus ou moins 2 écarts-types» est de 95% (plus exactement 0,9544). En termes de fréquence des observations on a coutume de dire que, *grosso modo,* 95% des observations doivent se situer dans cet intervalle. Cette propriété est d'ailleurs souvent étendue de façon tout à fait abusive à tout type de loi.
La troisième relation montre qu'il n'y a pratiquement aucune chance de trouver une observation au-delà de 3 écarts-types de la moyenne.

4.2.5 La loi lognormale $L\mathcal{N}(\mu, \sigma^2)$

Cette loi fournit souvent un bon modèle pour les variables strictement positives ayant une distribution asymétrique avec allongement vers les valeurs élevées, en particulier dans les domaines biologique (poids des personnes, par exemple), économique (distribution des revenus) et physique.

Soit X une v.a. à valeurs strictement positives, on dit qu'elle suit une *loi lognormale* de paramètres μ et σ^2, notée $L\mathcal{N}(\mu, \sigma^2)$, si $\ln X \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$.

Sa densité, peu utilisée car on préfère généralement se ramener à l'échelle logarithmique, peut être déduite de celle de la loi de Gauss par la transformation exponentielle selon la méthode du changement de variable exposée en section 1.6. Posons $Y = \ln X$, on a :

$$F(x) = P(X \leq x) = P(e^Y \leq x) = P(Y \leq \ln x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right)\ .$$

En dérivant, on obtient aisément :

$$f(x) = \begin{cases} \dfrac{1}{x\sqrt{2\pi}\sigma} exp\{-\dfrac{1}{2\sigma^2}(\ln x - \mu)^2\} & \text{si } x > 0\\ 0 & \text{si } x \leq 0 \end{cases}\ .$$

Comme $P(X \leq x) = P(\log X \leq \log x)$ les quantiles restent en correspondance par la transformation exponentielle et, par exemple, la médiane est e^{μ}. Il n'en va pas ainsi des moments qui peuvent toutefois se déduire directement de la fonction génératrice de la loi de Gauss $\mathcal{N}(\mu, \sigma^2)$. En effet, en posant encore $Y = \ln X$, on a :

$$E(X^k) = E(e^{kY}) = \Psi_Y(k) = e^{k\mu + \frac{k^2\sigma^2}{2}},$$

soit, en prenant $k = 1$ puis $k = 2$:

$$\begin{aligned} E(X) &= e^{\mu + \frac{1}{2}\sigma^2} \\ E(X^2) &= e^{2\mu + 2\sigma^2} \\ \text{d'où } : V(X) &= e^{2\mu + \sigma^2}(e^{\sigma^2} - 1)\,. \end{aligned}$$

.

4.2.6 La loi de Pareto

Cette loi a été introduite pour modéliser la **distribution de revenus** supérieurs à un seuil donné, puis s'est avérée utile pour d'autres phénomènes (par exemple la distribution de la taille de grains de sable passés au travers d'un tamis). Elle a deux paramètres strictement positifs : le paramètre de seuil a et un paramètre de forme θ. La fonction de répartition et la fonction de densité sont :

$$F(x) = \begin{cases} 1 - \left(\dfrac{a}{x}\right)^{\theta} & \text{si } x \geq a \\ 0 & \text{si } x < a \end{cases} \quad \text{et} \quad f(x) = \begin{cases} \dfrac{\theta}{a}\left(\dfrac{a}{x}\right)^{\theta+1} & \text{si } x \geq a \\ 0 & \text{si } x < a \end{cases} .$$

La densité étant une puissance de x, on calcule aisément (voir exercices du chapitre 2) :

$$\begin{aligned} E(X) &= \frac{\theta a}{\theta - 1} \ \text{(n'existe que si } \theta > 1)\ , \\ V(X) &= \frac{\theta a^2}{(\theta - 1)^2(\theta - 2)} \ \text{(n'existe que si } \theta > 2). \end{aligned}$$

Sa fonction génératrice des moments n'existe pas (sa fonction caractéristique - voir note 2.4 - ne s'exprime pas par des fonctions usuelles).

4.2.7 La loi de Weibull $W(\lambda, \alpha)$

Cette loi généralise la loi exponentielle pour modéliser des **durées de vie.** Elle intervient également dans les problèmes dits de valeurs extrêmes (par exemple l'occurrence de crues exceptionnelles d'une rivière). La fonction de

répartition et la fonction de densité de cette loi, notée $W(\lambda, \alpha)$ où λ et α sont deux paramètres strictement positifs, sont :

$$F(x) = \begin{cases} 1 - e^{-\lambda x^\alpha} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \quad \text{et} \quad f(x) = \begin{cases} \alpha\lambda x^{\alpha-1} e^{-\lambda x^\alpha} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases} .$$

Quand $\alpha = 1$ on a la loi $\mathcal{E}(\lambda)$, quand $\alpha < 1$ la densité décroît depuis $+\infty$, quand $\alpha > 1$ elle admet un maximum (*mode* de la loi) au point $[\frac{1}{\lambda}(\frac{\alpha-1}{\alpha})]^{1/\alpha}$.

On montre que :

$$E(X) = \frac{\Gamma(1 + \frac{1}{\alpha})}{\lambda^{1/\alpha}} \qquad , \qquad V(X) = \frac{\Gamma(1 + \frac{2}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha})}{\lambda^{2/\alpha}}$$

où Γ est la fonction gamma d'Euler (voir section 4.2.3).

Montrons quelques particularités utiles pour la modélisation de durées de vie.

Proposition 4.5 *Si $X \rightsquigarrow \mathcal{E}(\lambda)$ alors $X^{1/\alpha}$ suit une loi $W(\lambda, \alpha)$.*

Cette proposition est évidente par le principe du changement de variable exposé en section 1.6. Ainsi pour $\alpha > 1$ cela revient à une contraction de l'échelle du temps et donc à introduire un effet d'usure. Considérons en effet, comme nous l'avons fait pour la loi exponentielle, la probabilité qu'un système «survive» un temps h fixé ($h > 0$) au-delà du temps t et étudions cette probabilité comme une fonction ρ de t. On a :

$$\rho(t) = P(X > t + h | X > t) = \frac{P(X > t + h)}{P(X > t)} = e^{-\lambda[(t+h)^\alpha - t^\alpha]}.$$

La fonction $(t + h)^\alpha - t^\alpha$ étant croissante pour $\alpha > 1$ et décroissante pour $\alpha < 1$, la probabilité diminue avec le temps pour $\alpha > 1$ ce qui correspond bien à un phénomène d'usure. Au contraire pour $\alpha < 1$ on a une probabilité qui augmente (on peut penser ici à la durée de chômage où plus le temps s'écoule plus il est difficile d'en sortir).

4.2.8 La loi de Gumbel

C'est une autre loi de modélisation de valeurs extrêmes dont la fonction de répartition est :

$$F(x) = exp\left\{-e^{-\frac{x-\alpha}{\beta}}\right\} \quad , \quad x \in \mathbb{R} \quad (\beta > 0) .$$

On montre que sa moyenne est $\alpha + \gamma\beta$, où γ =0,577... est la constante d'Euler, et que sa variance est $\pi^2\beta^2/6$.

La valeur α correspond à son mode. Sa fonction génératrice des moments est $\Psi(t) = e^{\alpha t}\Gamma(1 - \beta t)$. Elle est liée à la loi limite du maximum d'une série de n observations quand n tend vers l'infini, pour une grande variété de lois.

4.2.9 La loi bêta $Beta(\alpha, \beta)$

Cette loi fournit un modèle pour les mesures comprises entre 0 et 1, en particulier pour des taux ou des proportions. Sa densité est :

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)} x^{\alpha}(1-x)^{\beta} & \text{si } x \in]0;1[\\ 0 & \text{si } x \notin]0;1[\end{cases}$$

avec $\alpha > -1$ et $\beta > -1$.

Pour $\alpha = \beta = 0$ on a la loi uniforme $\mathcal{U}[0,1]$. Pour α et β strictement positifs elle admet un mode en $x = \alpha/(\alpha+\beta)$.

Sachant que, pour tout $\alpha > -1$ et tout $\beta > -1$, on a :

$$\int_0^1 x^{\alpha}(1-x)^{\beta} dx = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)},$$

on calcule aisément, pour $X \rightsquigarrow Beta(\alpha, \beta)$:

$$E(X) = \frac{\alpha+1}{\alpha+\beta+2}$$
$$V(X) = \frac{(\alpha+1)(\beta+1)}{(\alpha+\beta+2)^2(\alpha+\beta+3)}.$$

4.3 Génération de nombres issus d'une loi donnée

Il n'est pas toujours possible d'étudier de façon analytique le comportement de modèles, d'estimateurs ou de statistiques de tests en raison de leur complexité. Dans ce cas on recourt à des **simulations** d'échantillons pour suppléer à l'absence d'éléments théoriques et nous nous réfèrerons parfois, dans les chapitres ultérieurs, à des résultats obtenus de cette façon. Cette approche constitue l'essence de la *méthode de Monte-Carlo.*

Tous les logiciels offrant des possibilités de calcul disposent d'un générateur de «nombres au hasard» (fonction RANDOM, ALEA, etc., voir section 4.2.1) qui correspondent à des observations issues d'une loi $\mathcal{U}[0\,,1]$ ou que l'on peut considérer comme telles. Car, en réalité, ces nombres que l'on qualifie plutôt de *pseudo-aléatoires*, sont engendrés par un mécanisme purement déterministe.

Supposons que l'on veuille générer des réalisations d'une loi continue de fonction de répartition F strictement croissante et que l'on dispose de la fonction inverse F^{-1}, soit de façon analytique, soit de façon numérique (de nombreux logiciels statistiques ou autres proposent, par exemple, les fonctions

«Gauss-inverse», «exponentielle-inverse», etc.). Étant donné une variable aléatoire U de loi $\mathcal{U}[0\,,1]$, considérons la fonction $X = F^{-1}(U)$ et déterminons sa fonction de répartition en appliquant la méthode de la section 1.6. On a :

$$\begin{aligned} P(X \leq x) &= P(F^{-1}(U) \leq x) \\ &= P(U \leq F(x)) = F(x) \end{aligned}$$

puisque, pour la loi uniforme, $P(U \leq u) = u$. Donc X suit la loi F.

Ainsi, à partir d'une suite de nombres au hasard $u_1, u_2, \cdots, u_n$ on peut obtenir une suite de nombres $x_1, x_2, \cdots, x_n$ issus de la loi F, par la transformation $x_i = F^{-1}(u_i)$.

Pour la loi $\mathcal{E}(\lambda)$, par exemple, $F(x) = 1 - e^{-\lambda x}$ et la fonction inverse est explicite : $F^{-1}(x) = -\frac{1}{\lambda}\ln(1-x)$. On utilisera alors la transformation $x_i = -\frac{1}{\lambda}\ln(1-u_i)$.

Pour une loi discrète la méthode ci-dessus n'est pas applicable du fait que F, étant une fonction en escalier, n'est pas inversible. Dans le cas où le nombre de valeurs possibles $a_1 < a_2 < \cdots < a_k < \cdots < a_r$ est restreint on peut l'adapter de la façon suivante :

$$\begin{aligned} &\text{si} \quad u_i \in [0, F(a_1)[\quad \text{alors générer} \quad x_i = a_1 \\ &\qquad \cdots \\ &\text{si} \quad u_i \in [F(a_{k-1}), F(a_k)[\quad \text{alors générer} \quad x_i = a_k \\ &\qquad \cdots \\ &\text{si} \quad u_i \in [F(a_{r-1}), 1] \quad \text{alors générer} \quad x_i = a_r\,, \end{aligned}$$

le choix d'ouvrir ou de fermer chaque intervalle d'un côté ou de l'autre n'ayant pas d'importance si les u_i sont générés avec suffisamment de décimales.

En particulier on peut produire un processus de Bernoulli de paramètre p en donnant la valeur 1 si $u_i < p$ et 0 si $u_i \geq p$.

Il existe toutefois des méthodes de génération adaptées et plus efficaces pour chaque loi que l'on trouvera dans les ouvrages consacrés spécifiquement à la simulation.

4.4 Exercices

Exercice 4.1 En transformant linéairement la v.a. de la marche aléatoire (voir exercices du chapitre 3) en une v.a. de Bernoulli, établir la loi de cette marche aléatoire après n pas.

Exercice 4.2 Montrer directement que la fonction génératrice des moments $\Psi(t)$ de la loi binomiale négative est $p^r/[1-(1-p)e^t]^r$ et qu'elle est définie pour $t < -\ln(1-p)$.

Aide : substituer à $(1-p)^x$ l'expression $[(1-p)e^t]^x$.
En écrivant que la somme des termes de la fonction de probabilité vaut 1 et en dérivant terme par terme par rapport à p, déduire l'expression de la moyenne de la loi.

Exercice 4.3 * (Approche historique de Moivre mettant en évidence la loi de Gauss) Soit $X \rightsquigarrow \mathcal{B}(n,p)$, montrer que pour $n \to \infty$, p restant fixe, la probabilité $P(X) = x$ est équivalente à la fonction de densité en x de la loi de $U \rightsquigarrow \mathcal{N}(np, np(1-p))$ ou encore à $P(x - \frac{1}{2} < U < x + \frac{1}{2})$. On admettra intuitivement que les valeurs d'intérêt pour x (à probabilités non négligeables) tendent vers l'infini quand $n \to \infty$ et que, pour ces valeurs, $\frac{x}{n} \to p$.
On utilisera pour la démonstration la formule de Stirling[2] :

$$n! = \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} (1 + o(1)) \text{ soit } n! \sim \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}$$

Exercice 4.4 Soit $X \rightsquigarrow \mathcal{B}(n,p)$. Montrer que si $n \to \infty$ et $p \to 0$ de façon que np reste constant, alors $P(X = x)$ tend vers la probabilité correspondante de la loi de Poisson de paramètre np.

Aide : on admettra que $\frac{n!}{(n-x)!n^x} \to 1$ quand $n \to \infty$.

Exercice 4.5 Soit $X \rightsquigarrow \mathcal{H}(N,M,n)$. Montrer que, quand $N \to \infty$ et $\frac{M}{N} \to p$ (non nul), $P(X = x)$ tend vers la probabilité correspondante de la loi $\mathcal{B}(n,p)$.

Aide : comme pour l'exercice précédent.

Exercice 4.6 Soient X_1 et X_2 deux v.a. indépendantes de Poisson de paramètres respectifs λ_1 et λ_2. Montrer que la loi conditionnelle de X_1 sachant $X_1 + X_2 = n$ est une loi binomiale.

Exercice 4.7 Soit $X \rightsquigarrow \mathcal{G}(p)$. Déterminer $P(X > n)$ et montrer que la probabilité $P(X > n+k \mid X > n)$ est indépendante de n. [Note : Ceci est à rapprocher de la propriété analogue de la loi $\mathcal{E}(\lambda)$. La loi $\mathcal{G}(p)$ peut modéliser la durée de vie d'un système sans usure, en temps discret d'intervalles réguliers].

Exercice 4.8 Soit $X \rightsquigarrow \mathcal{U}[0,1]$, montrer que $Y = (b-a)X + a$ suit une loi $\mathcal{U}[a,b]$.

Exercice 4.9 Montrer que la fonction génératrice des moments de la loi $\Gamma(r,\lambda)$, pour $r > 0$ non nécessairement entier, est $\Psi(t) = [\lambda/(\lambda - t)]^r$. Pour quelles valeurs de t est-elle définie ? En déduire sa moyenne et sa variance. [Rappel sur la fonction gamma d'Euler : $\Gamma(r) = \int_0^{+\infty} x^{r-1} e^{-x} dx$ avec $r > 0$].

Exercice 4.10 En s'appuyant sur un processus de Poisson sous-jacent, déterminer pour r entier la fonction de répartition de la loi $\Gamma(r,\lambda)$. En déduire sa densité.

[2]Dans l'expression de cette formule le terme $o(1)$ indique une fonction qui devient négligeable devant 1 (donc qui tend vers zéro) quand n tend vers l'infini.

Exercice 4.11 Soit $X \rightsquigarrow \Gamma(r, \lambda)$, montrer que rX suit une loi $\Gamma(r, \frac{\lambda}{r})$. Montrer que λX suit une loi $\Gamma(r, 1)$.

Exercice 4.12 Montrer que si X suit une loi de Pareto dont le paramètre de seuil a est égal à 1, alors $\ln X$ suit une loi exponentielle.

Exercice 4.13 Le temps moyen de service à un distributeur de billets est de 30 secondes. Vous arrivez et trouvez cinq personnes en attente (la première venant juste d'accéder au guichet). Quelle est la probabilité que vous attendiez moins de 30 secondes (on supposera être en présence d'un processus de Poisson) ?

Aide : on utilisera le deuxième résultat de l'exercice 4.11 et on établira une relation de récurrence pour $I_n = \int_0^1 x^n e^{-x} dx$ en intégrant par parties.

Exercice 4.14 Pour un projet de construction d'un immeuble de 20 logements, on étudie la capacité nécessaire du parking. On note X la variable «nombre de voitures d'un ménage». Pour tout ménage on admet que la probabilité d'avoir une voiture est 0,70 et celle d'avoir 2 voitures est de 0,30 (on néglige toute autre possibilité). On supposera l'indépendance du nombre de voitures entre les ménages.
On pose $Y = X - 1$. Quelle est la loi de Y ?
Quelle est la loi de la somme de 20 variables i.i.d. de même loi que Y ? En déduire la probabilité qu'un parking de 29 places soit suffisant pour les 20 ménages.

Exercice 4.15 Grâce à une importante étude épidémiologique on constate que la distribution des poids des individus dans une population adulte donnée peut être convenablement modélisée par une loi lognormale. Considérant que le poids moyen est de 70 kg et que l'écart-type des poids est de 12 kg résoudre les deux équations permettant de déterminer les valeurs des paramètres μ et σ^2 de la loi lognormale.

Chapitre 5

Lois fondamentales de l'échantillonnage

5.1 Phénomènes et échantillons aléatoires

Nous entrons maintenant véritablement dans le domaine de la statistique en nous penchant sur l'étude d'**observations répétées** issues d'un certain phénomène de nature aléatoire.

Schématiquement, on peut distinguer deux classes de phénomènes aléatoires. D'une part l'aléatoire peut être provoqué expérimentalement comme, par exemple, dans les jeux de hasard ou dans les mécanismes de tirage au sort «d'individus» dans des «populations[1]» finies pour les sondages, pour le contrôle de qualité, etc.(voir section 3.7). Dans ce contexte expérimental la notion d'expérience aléatoire, point de départ de la modélisation probabiliste, a un sens tout à fait réel.

D'autre part, on peut aussi recourir à une modélisation aléatoire lorsqu'on est incapable de prévoir avec exactitude les réalisations d'un phénomène. Le caractère aléatoire est simplement attribué au phénomène pour refléter l'incertitude de l'observateur par rapport à un ensemble de résultats possibles, par exemple le nombre d'appels parvenant à un standard téléphonique dans une unité de temps, la durée de vie d'un appareil, etc. Il n'y a pas ici d'expérience aléatoire à proprement parler. Toutefois il est nécessaire, pour l'approche statistique, de pouvoir observer le phénomène de façon répétée afin de constituer des échantillons.

[1]Nous mettons ces termes entre guillemets car ils sont à prendre dans un sens large et non uniquement par référence à des populations humaines. A proprement parler les «individus» sont des *unités statistiques* qui peuvent être les entreprises d'un secteur d'activité, les arbres d'une forêt, les pièces d'un lot de production, etc. La population est aussi appelée *univers*.

Définition 5.1 *On appelle* ***échantillon aléatoire de taille*** n *(en bref* n*-échantillon) une suite de* n *variables aléatoires indépendantes et de même loi (ou v.a. i.i.d). Cette loi est appelée la* ***loi mère*** *de l'échantillon.*

Cette définition appelle quelques remarques.

- Mathématiquement la notion d'échantillon aléatoire est identique à celle de v.a. i.i.d., et l'usage de ce terme ne se justifie qu'en raison du contexte de l'échantillonnage. Sauf mention contraire, quand on parlera d'échantillon dans cet ouvrage, il s'agira implicitement d'une suite de v.a. i.i.d.
- Il sera commode d'associer à la loi mère un symbole de v.a., par exemple X, le n-échantillon étant alors désigné par $X_1, X_2, ..., X_n$. Ainsi on peut écrire que pour tout $i = 1, ..., n$, $E(X_i) = E(X)$ qui représente la moyenne de la loi mère.
- On parle souvent, en lieu et place de la loi mère, de *la distribution de la population* - voire même simplement de la population - en référence à un sondage. Certes ce terme est abusif car, d'une part on y confond les individus et les valeurs numériques observables sur ces individus et, d'autre part, il n'existe pas nécessairement une population réelle (quelle est la population des appels à un standard, des produits d'un certain type manufacturés par une entreprise ?). Toutefois il nous arrivera de recourir à ce terme comme s'il existait une sorte de *population virtuelle* dont les observations seraient issues comme par un tirage au hasard.
- Le statut de v.a. i.i.d. exige que le phénomène soit invariant au cours des observations successives et que ces observations n'exercent aucune influence entre elles. Il s'agit bien souvent d'une profession de foi, ces conditions n'étant généralement pas rigoureusement vérifiables, ni rigoureusement vérifiées.
- Pour ce qui est des notations on distinguera la notion d'**échantillon aléatoire** $X_1, X_2, ..., X_n$ dont on peut dire qu'elle se réfère à des résultats potentiels avant expérience ou *a priori,* de celle d'**échantillon réalisé** $x_1, x_2, ..., x_n$ correspondant aux valeurs observées après expérience ou *a posteriori.*

L'objectif de ce chapitre est d'étudier certaines caractéristiques de l'échantillon aléatoire, essentiellement sa moyenne et sa variance, en relation avec celles de la loi mère. A priori (au sens de la remarque précédente) une telle caractéristique est une v.a. qui prend le nom de «statistique» dans le contexte de l'échantillonnage, selon la définition suivante.

Définition 5.2 *Soit* $X_1, X_2, ..., X_n$ *un n-échantillon, on appelle* ***statistique*** *toute v.a.* $T_n = h(X_1, X_2, ..., X_n)$*, fonction de* $X_1, X_2, ..., X_n$*.*

On peut concrétiser la loi d'une statistique (donc d'une caractéristique, telle la moyenne de l'échantillon) en imaginant une simulation en très grand nombre

d'échantillons de taille n, en calculant pour chacun d'eux la valeur prise par la statistique et en étudiant la distribution de ces valeurs. De façon imagée on peut dire qu'il s'agit de la *distribution d'échantillonnage* de la statistique sur «l'univers» de tous les échantillons possibles. Notons qu'une statistique peut être une fonction à valeurs dans $\mathbb{R}$, $\mathbb{R}^2$ ou $\mathbb{R}^p$. En particulier les moments empiriques ci-après sont à valeurs dans $\mathbb{R}$. Les définitions qui suivent se rapportent toutes à un échantillon aléatoire noté $X_1, X_2, ..., X_n$.

5.2 Moyenne, variance, moments empiriques

Définition 5.3 *On appelle* ***moyenne de l'échantillon*** *ou* ***moyenne empirique*** *la statistique, notée $\bar{X}$, définie par :*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Définition 5.4 *On appelle* ***variance empirique*** *la statistique, notée $\widetilde{S}^2$, définie par :*

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

Nous commençons maintenant à établir certaines relations entre les lois de ces statistiques et la loi mère.

Proposition 5.1 *Soit μ et σ^2, respectivement la moyenne et la variance de la loi mère. On a :*

$$E(\bar{X}) = \mu \ , \qquad V(\bar{X}) = \frac{\sigma^2}{n} \, .$$

En effet :

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

Puis, en raison de l'indépendance des X_i :

$$V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Proposition 5.2 *La moyenne de la loi de la variance empirique est :*

$$E(\widetilde{S}^2) = \frac{n-1}{n}\sigma^2 \, .$$

En effet :

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

D'où :

$$E(\widetilde{S}^2) = \frac{1}{n}\sum_{i=1}^{n}V(X_i) - V(\bar{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Quand on abordera l'estimation ponctuelle (chapitre 6) on dira que $\widetilde{S}^2$ est un estimateur biaisé de σ^2. Par anticipation définissons l'estimateur $S^2 = \frac{n}{n-1}\widetilde{S}^2$ qui est sans biais pour σ^2, c'est-à-dire tel que $E(S^2) = \sigma^2$.

Définition 5.5 *On appelle* ***variance de l'échantillon*** *la statistique*

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

Dorénavant on étudiera S^2 plutôt que $\widetilde{S}^2$ à laquelle on pourra éventuellement se référer en conservant le terme de variance empirique.

En prenant la racine carrée de S^2 (respectivement de $\widetilde{S}^2$) on définit *l'écart-type* S de l'échantillon (respectivement, l'écart-type empirique $\widetilde{S}$).

Cas particulier : loi mère gaussienne

Si la loi mère est $\mathcal{N}(\mu,\sigma^2)$ alors $\bar{X}$ est gaussienne, en tant que combinaison linéaire de gaussiennes indépendantes (voir proposition 4.4). Par conséquent :

$$\bar{X} \rightsquigarrow \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

La loi de S^2 sera vue dans la section 5.3. Par ailleurs, nous admettrons la proposition suivante.

Proposition 5.3 *Si la loi mère est gaussienne,* $\bar{X}$ *et* S^2 *sont des v.a. indépendantes.*

Note 5.1 : D'une façon générale, la loi de $\bar{X}$ (et a fortiori celle de S^2), pour une loi mère quelconque, n'est pas facile à identifier. Des cas particuliers seront vus en exercice (loi mère exponentielle, loi mère de Cauchy). Si la loi mère est de Bernoulli $\mathcal{B}(p)$ ou de Poisson $\mathcal{P}(\lambda)$ on peut déduire la loi de $\bar{X}$ à partir de celle de la somme $\sum_{i=1}^{n} X_i = n\bar{X}$ qui est, respectivement, une loi binomiale $\mathcal{B}(n,p)$ ou une loi de Poisson $\mathcal{P}(n\lambda)$ selon les propriétés de ces lois vues au chapitre 4.

Définition 5.6 *On appelle* ***moment empirique*** *d'ordre r, noté M_r, la statistique*

$$M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

Définition 5.7 *On appelle* ***moment centré empirique*** *d'ordre r, noté M'_r, la statistique*

$$M'_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r.$$

Proposition 5.4 *Si la loi mère admet un moment μ_r d'ordre r (voir définition 2.2) alors :*

$$E(M_r) = \mu_r.$$

Ceci découle directement du fait que, pour tout $i = 1, ..., n$, $E(X_i^r) = E(X^r) = \mu_r$ (où X est le symbole de v.a. associé à la loi mère). On verra en fait, plus loin, qu'un moment empirique est un «estimateur» naturel du moment de même ordre de la loi (appelé parfois, par contraste, *moment théorique*). En revanche, comme on l'a vu pour la variance empirique $\widetilde{S}^2$ qui correspond au moment empirique centré d'ordre 2, $E(M'_r)$ n'est pas nécessairement égal à μ'_r, le moment centré théorique de même ordre. Ceci résulte du fait que le centrage est effectué avec la moyenne de l'échantillon $\bar{X}$ et non pas avec la vraie moyenne μ de sa loi. Les moments centrés empiriques s'expriment, par développement des $(X_i - \bar{X})^r$, en fonction des moments empiriques simples de la même façon que le font les moments théoriques entre eux, puisqu'il s'agit alors de développer $(X - \mu)^r$.

En particulier, on a la **formule de décentrage de la variance empirique** :

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

qui fait le pendant de celle de la section 2.3 : $V(X) = E(X^2) - \mu^2$.

Note 5.2 : En considérant un n-échantillon de couples d'observations $(X_1, Y_1), ..., (X_n, Y_n)$ on peut définir des moments croisés empiriques d'ordres p et q quelconques et leurs correspondants centrés :

$$\frac{1}{n}\sum_{i=1}^{n} X_i^p Y_i^q \quad \text{et} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^p (Y_i - \bar{Y})^q.$$

Pour $p = q = 1$ le moment centré est la *covariance empirique* utilisée en statistique descriptive :

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})\,.$$

Comme pour la variance on introduit le facteur $\frac{1}{n-1}$ au lieu de $\frac{1}{n}$ pour éliminer le biais vis-à-vis de la covariance théorique (voir définition 3.5).

Par analogie avec la définition 3.6 de la corrélation linéaire on définit la *corrélation linéaire empirique* en divisant la covariance empirique par le produit des écarts-types empiriques des v.a. X et Y, soit après simplification :

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})\,(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}\,,$$

formule bien connue en statistique descriptive.

Nous abordons maintenant trois lois omniprésentes en statistique car liées aux distributions d'échantillonnage de moyennes et de variances dans le cas gaussien.

5.3 Loi du Khi-deux

Définition 5.8 *Soit $Z_1, Z_2, ..., Z_\nu$ une suite de variables aléatoires i.i.d. de loi $\mathcal{N}(0\,;1)$. Alors la v.a. $\sum_{i=1}^{\nu} Z_i^2$ suit une loi appelée* ***loi du Khi-deux à ν degrés de liberté****, notée $\chi^2(\nu)$.*

Proposition 5.5 *La densité de la loi du Khi-deux à ν degrés de liberté est :*

$$f_\nu(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}} \quad \textit{pour } x > 0 \;\; (0 \;\; \textit{sinon}).$$

Démonstration : calculons la fonction génératrice de Z^2 où $Z \rightsquigarrow \mathcal{N}(0\,;1)$. On a :

$$\begin{aligned}\Psi_{Z^2}(t) &= E(e^{tZ^2}) = \int_{-\infty}^{+\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}(1-2t)z^2} dz \\ &= \frac{1}{\sqrt{1-2t}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}u^2} du \text{ en posant } u = \sqrt{1-2t}\, z \\ &= \frac{1}{\sqrt{1-2t}}\end{aligned}$$

qui est définie pour $t < \frac{1}{2}$.

Pour la somme des Z_i^2 indépendantes on a donc (voir proposition 3.12) :

$$\Psi_{\sum_{i=1}^{\nu} Z_i^2}(t) = (\frac{1}{1-2t})^{\frac{\nu}{2}}$$

qui n'est autre que la fonction génératrice d'une loi $\Gamma(\frac{\nu}{2}\ , \frac{1}{2})$ vue en section 4.2.3, dont la densité est bien celle de la proposition ci-dessus. On voit donc, au passage, qu'une loi du Khi-deux est un cas particulier de loi gamma. □

Proposition 5.6 *La moyenne de la loi $\chi^2(\nu)$ est égale au nombre de degrés de liberté ν, sa variance est 2ν.*

Repartons de la définition de la loi $\chi^2(\nu)$. Comme $Z_i \rightsquigarrow \mathcal{N}(0\,;1)$ on a :

$$\begin{aligned}&E(Z_i^2) = V(Z_i) = 1 \quad \text{d'où} \quad E(\sum_{i=1}^{\nu} Z_i^2) = \nu \\ &V(Z_i^2) = E(Z_i^4) - (E(Z_i^2))^2 = \mu_4 - 1.\end{aligned}$$

Or d'après un résultat de la section 4.2.4, $\mu_4 = 3$ d'où $V(Z_i^2) = 2$ et $V(\sum_{i=1}^{n} Z_i^2) = 2\nu$.

Proposition 5.7 *Si $T_1 \rightsquigarrow \chi^2(\nu_1)$, $T_2 \rightsquigarrow \chi^2(\nu_2)$, T_1 et T_2 indépendantes, alors $T_1 + T_2 \rightsquigarrow \chi^2(\nu_1 + \nu_2)$.*

Cette proposition est évidente de par la définition de la loi du Khi-deux. Nous revenons maintenant sur la loi de S^2 dans le cas d'un échantillon de loi mère gaussienne.

Théorème 5.1 *Soit un n-échantillon $X_1, X_2, ..., X_n$ de loi $\mathcal{N}(\mu, \sigma^2)$ on a :*

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1).$$

Démonstration : en reprenant les développements qui suivent l'énoncé de la proposition 5.2, on établit que :

$$\sum_{i=1}^{n}(X_i-\mu)^2 \quad = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2,$$

$$\text{soit} \qquad \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \quad = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} + \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

Les deux termes de droite sont, respectivement, $\frac{(n-1)S^2}{\sigma^2}$ et le carré de $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ qui est une gaussienne centrée-réduite. Ces termes aléatoires étant indépendants (comme fonctions de S^2 et de $\bar{X}$, voir proposition 5.3) et les v.a. $\frac{X_i-\mu}{\sigma}$ étant indépendantes de loi $\mathcal{N}(0\,;1)$ on a, en termes de fonctions génératrices :

$$\left(\frac{1}{1-2t}\right)^{n/2} = \Psi_{\frac{(n-1)S^2}{\sigma}}(t).\left(\frac{1}{1-2t}\right)^{1/2} (\text{si } t<\frac{1}{2}).$$

Finalement :

$$\Psi_{\frac{(n-1)S^2}{\sigma}}(t) = \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}},$$

ce qui prouve le théorème. □

Sachant que l'espérance d'une loi $\chi^2(n-1)$ est $n-1$ et sa variance $2(n-1)$, on voit que :

$$E(S^2)=\sigma^2 \quad \text{et} \quad V(S^2)=\frac{2\sigma^4}{n-1}.$$

En fait la loi du Khi-deux a un usage beaucoup plus vaste en statistique notamment dans la théorie des tests comme nous le verrons au chapitre 9. Toutes ses applications reposent sur des sommes de carrés de termes gaussiens ou approximativement gaussiens. Notons, finalement, que la fonction de répartition de la loi (ou plutôt de la famille de lois) du Khi-deux ne s'explicite pas et que l'on doit recourir à des tables ou à une fonction *ad hoc* dans les logiciels statistiques pour le calcul de probabilités.

5.4 Loi de Student

Définition 5.9 *Soit Z et Q deux v.a. indépendantes telles que $Z \rightsquigarrow \mathcal{N}(0\,;1)$ et $Q \rightsquigarrow \chi^2(\nu)$. Alors la v.a.*

$$T=\frac{Z}{\sqrt{\frac{Q}{\nu}}}$$

suit une loi appelée ***loi de Student à ν degrés de liberté****, notée $t(\nu)$.*

Proposition 5.8 *La densité de la loi de Student à ν degrés de liberté est :*

$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} , \quad x \in \mathbb{R}\,.$$

Ce résultat, que nous ne démontrons pas, est dû à W.S. Gosset en 1908, qui prit le pseudonyme de Student. Ni la fonction de répartition, ni la fonction génératrice ne s'explicitent. Il existe donc des tables de la fonction de répartition ou une fonction *ad hoc* dans les logiciels statistiques. On admettra encore la proposition suivante.

Proposition 5.9 *Soit $T \rightsquigarrow t(\nu)$ alors $E(T) = 0$ si $\nu \geq 2$ et $V(T) = \frac{\nu}{\nu-2}$ si $\nu \geq 3$.*

Le fait que la moyenne est nulle est évident puisque la densité est une fonction paire. On notera que la variance vaut 3 dès qu'elle est définie ($\nu = 3$) et tend vers 1 quand $\nu \to +\infty$. Pour être plus précis, l'allure de la loi de Student est similaire à celle d'une loi de Gauss centrée-réduite avec un étalement un peu plus fort, cette différence s'estompant rapidement lorsque ν s'accroît et devenant négligeable pour $\nu > 200$. Ceci s'explique, en fait, par sa définition même mettant en jeu une v.a. $\mathcal{N}(0\,;1)$ au numérateur et une v.a. qui converge en probabilité (voir cette notion en section 5.8.1) vers 1, au dénominateur.

Pour $\nu = 1$ la loi est la *loi de Cauchy*. Selon la définition 5.9, c'est la loi du rapport de deux gaussiennes centrées et réduites indépendantes (cette définition impose en fait de prendre la valeur absolue de la variable du dénominateur, mais cette restriction peut être levée). Sa moyenne n'existe pas en raison de ses "queues de distribution" pesantes (voir exemple 2.1), ce qui lui confère certaines particularités, par exemple que la loi des grands nombres (voir section 5.8.2) ne s'applique pas (voir exercices).

Théorème 5.2 *(Application fondamentale) Soit $X_1, X_2, ..., X_n$ un n-échantillon de loi mère $\mathcal{N}(\mu, \sigma^2)$. Alors :*

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \rightsquigarrow t(n-1).$$

La démonstration est immédiate en prenant, avec les notations utilisées pour la définition 5.9,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{et} \quad Q = \frac{(n-1)S^2}{\sigma^2}.$$

Ce résultat, qui a motivé en réalité les travaux de Gosset, met en évidence la modification apportée à la loi $\mathcal{N}(0\,;1)$ de la v.a. Z ci-dessus, lorsqu'on substitue à l'écart-type théorique σ de la loi mère, l'écart-type de l'échantillon S. On

comprend au passage, qu'introduisant un terme aléatoire supplémentaire, on provoque un étalement plus grand. En fait les applications de la loi de Student se rencontrent souvent en statistique dès lors qu'on est appelé à remplacer σ, en général inconnu, par son «estimateur» naturel S.

5.5 Loi de Fisher-Snedecor

Définition 5.10 *Soit U et V deux v.a. indépendantes telles que $U \rightsquigarrow \chi^2(\nu_1)$ et $V \rightsquigarrow \chi^2(\nu_2)$. Alors la v.a.*

$$F = \frac{U/\nu_1}{V/\nu_2}$$

suit une ***loi de Fisher-Snedecor à ν_1 degrés de liberté au numérateur et ν_2 degrés de liberté au dénominateur****, notée $F(\nu_1, \nu_2)$. En bref on l'appellera loi de Fisher.*

Proposition 5.10 *La densité de la loi $F(\nu_1, \nu_2)$ est :*

$$f_{\nu_1,\nu_2}(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{x^{\frac{\nu_1-2}{2}}}{(1+\frac{\nu_1}{\nu_2}x)^{\frac{\nu_1+\nu_2}{2}}} \quad si\ x > 0 \ \ (0\ sinon)\ .$$

Si $\nu_2 \geq 3$ sa moyenne existe et est égale à $\frac{\nu_2}{\nu_2-2}$. Si $\nu_2 \geq 5$ sa variance existe et est égale à $\frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$.

Nous admettrons ces résultats sans démonstration, notant avec curiosité que la moyenne, quand elle existe, ne dépend que des degrés de liberté du dénominateur. La fonction de répartition (tout comme la fonction génératrice) n'étant pas explicite il existe des tables ou des fonctions *ad hoc* dans les logiciels. La proposition suivante permet une économie de tables.

Proposition 5.11 *Soit $H \rightsquigarrow F(\nu_1, \nu_2)$, alors $\frac{1}{H} \rightsquigarrow F(\nu_2, \nu_1)$.*

Cette proposition est évidente de par la définition même de la loi de Fisher.

Montrons qu'il suffit, grâce à cette propriété, de disposer des quantiles d'ordre supérieur à 0,50. Soit à calculer, par exemple, pour $H \rightsquigarrow F(\nu_1, \nu_2)$, le quantile d'ordre 0,05. On a :

$$\begin{aligned} P(H < h_{0,05}) &= 0{,}05 \\ P\left(\frac{1}{H} > \frac{1}{h_{0,05}}\right) &= 0{,}05 \\ P\left(\frac{1}{H} < \frac{1}{h_{0,05}}\right) &= 0{,}95\,. \end{aligned}$$

Il est donc égal à l'inverse du quantile 0,95 lu sur la loi $F(\nu_2, \nu_1)$. Plus généralement le quantile d'ordre α de la loi $F(\nu_1, \nu_2)$ est l'inverse du quantile d'ordre $1-\alpha$ de la loi $F(\nu_2, \nu_1)$.

Les applications de la loi de Fisher sont nombreuses en statistique dès lors que l'on veut étudier le rapport de deux sommes de carrés de termes gaussiens indépendants. L'application la plus directe concerne la loi du rapport des variances S_1^2/S_2^2 de deux échantillons indépendants de tailles respectives n_1 et n_2, issus de deux lois mères gaussiennes ayant une **même variance** σ^2. En effet :

$$\frac{(n_1-1)S_1^2}{\sigma^2} \rightsquigarrow \chi^2(n_1-1) \quad , \quad \frac{(n_2-1)S_2^2}{\sigma^2} \rightsquigarrow \chi^2(n_2-1),$$

d'où immédiatement :

$$\frac{S_1^2}{S_2^2} \rightsquigarrow F(n_1-1, n_2-1).$$

On remarquera encore que si $T \rightsquigarrow t(\nu)$ alors $T^2 \rightsquigarrow F(1\,, \nu)$.

5.6 Statistiques d'ordre

Cette notion est très utile dans une série de problèmes, notamment ceux de minima et de maxima (voir les exercices) que nous abordons tout d'abord. Comme précédemment nous considérons un échantillon aléatoire $X_1, X_2, ..., X_n$ dont la loi mère a pour fonction de répartition F.

Pour une série de nombres réels $(x_1, x_2, ..., x_n)$ notons $\max\{x_1, x_2, ..., x_n\}$ la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ qui lui associe le nombre maximal de cette série. On peut donc définir une v.a., notée $X_{(n)}$, fonction de $(X_1, X_2, ..., X_n)$ par :

$$X_{(n)} = \max\{X_1, X_2, ..., X_n\}.$$

La fonction de répartition de cette statistique se déduit aisément de F. En effet l'événement $(X_{(n)} \leq x)$ est équivalent à $(X_1 \leq x, X_2 \leq x, ..., X_n \leq x)$. Par conséquent :

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_1 \leq x\,,\, X_2 \leq x\,, ...,\, X_n \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x)...P(X_n \leq x) \quad \text{(indépendance)} \\ &= [F(x)]^n \quad \text{(même loi)}. \end{aligned}$$

De façon similaire on note $X_{(1)} = \min\{X_1, X_2, ..., X_n\}$ la fonction minimum et, en notant que l'événement $(X_{(1)} > x)$ est équivalent au fait que toutes les X_i sont supérieures à x, on a :

$$\begin{aligned} P(X_{(1)} > x) &= P(X_1 > x)P(X_2 > x)...P(X_n > x) \\ &= [1-F(x)]^n\,, \end{aligned}$$

d'où :

$$F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n .$$

Définition 5.11 *Soit h_k la fonction de $\mathbb{R}^n$ dans $\mathbb{R}$ qui à $(x_1, x_2, ..., x_n)$ fait correspondre la k-ième valeur parmi $x_1, x_2, ..., x_n$ lorsqu'on les range dans l'ordre croissant.*
On appelle alors ***statistique d'ordre*** *k, la v.a. notée $X_{(k)}$, définie par :*

$$X_{(k)} = h_k(X_1, X_2, ..., X_n).$$

Ceci généralise les notions de minimum ($k = 1$) et de maximum ($k = n$).

Proposition 5.12 *La fonction de répartition de $X_{(k)}$ est :*

$$F_{X_{(k)}}(x) = \sum_{j=k}^{n} \left(\begin{array}{c} n \\ j \end{array} \right) [F(x)]^j \, [1 - F(x)]^{n-j} .$$

Pour montrer cela il suffit de noter que l'événement $\{X_{(k)} \leq x\}$ est équivalent au fait qu'au moins k v.a. parmi $X_1, ..., X_n$ soient inférieures à k. Soit X la v.a. symbolisant la loi mère. Considérons l'expérience de Bernoulli avec pour «succès» l'événement $(X \leq x)$ dont la probabilité est $F(x)$. Le nombre de v.a. parmi $X_1, ..., X_n$ prenant une valeur inférieure ou égale à x est donc une v.a. de loi binomiale $\mathcal{B}(n, F(x))$. Pour obtenir la probabilité que ce nombre soit au moins égal à k on est amené à sommer les termes de cette binomiale de k à n.

Note 5.3 Considérons $X_{(i)}$ et $X_{(j)}$ avec $i < j$. La v.a. $U = X_{(j)} - X_{(i)}$ ne peut prendre que des valeurs positives et donc $P(U \geq 0) = 1$. De façon conventionnelle on écrira $P(X_{(j)} \geq X_{(i)}) = 1$ et même, de façon quelque peu rapide, $X_{(j)} \geq X_{(i)}$. Moyennant cette convention, il est possible, comme dans la plupart des ouvrages, de définir les statistiques d'ordre $X_{(1)}, X_{(2)}, ..., X_{(n)}$ par une permutation de $(X_1, X_2, ..., X_n)$ telle que $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$.

5.7 Fonction de répartition empirique

Nous abordons ici une variable aléatoire **fonctionnelle**, c'est-à-dire dont les réalisations sont en fait des fonctions. Nous nous contenterons de l'étudier en un point x fixé pour rester dans le cadre des variables aléatoires prenant leurs valeurs dans $\mathbb{R}$. Au chapitre 7 traitant des estimateurs fonctionnels on verra l'intérêt de la fonction de répartition empirique en tant qu'estimateur de F au même titre que $\bar{X}$ est un estimateur de μ, par exemple.

Définition 5.12 *Pour tout $x \in \mathbb{R}$, on appelle valeur de la* ***fonction de répartition empirique*** *en x, la statistique, notée $F_n(x)$, définie par :*

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

où $I_{(-\infty,x]}$ est la fonction indicatrice de l'intervalle $(-\infty, x]$, à savoir $I_{(-\infty,x]}(u) = 1$ si $u \in (-\infty, x]$ et 0 sinon.

En d'autres termes $F_n(x)$ est la v.a «proportion» des n observations $X_1, X_2, ..., X_n$ prenant une valeur inférieure ou égale à x. Chaque X_i ayant une probabilité $F(x)$ d'être inférieure ou égale à x, $nF_n(x)$ suit une loi binomiale $\mathcal{B}(n, F(x))$. En conséquence $F_n(x)$ est une v.a discrète prenant les valeurs $\frac{k}{n}$, où $k = 0, 1, ..., n$, avec probabilités :

$$P(F_n(x) = \frac{k}{n}) = P(nF_n(x) = k) = \left(\begin{array}{c} n \\ k \end{array} \right) [F(x)]^k \, [1 - F(x)]^{n-k} .$$

Note 5.4 A l'issue de l'expérience d'échantillonnage, soit $x_1, x_2, ..., x_n$ la réalisation du n-échantillon $X_1, X_2, ..., X_n$. La fonction de répartition empirique se réalise comme une fonction réelle définie sur tout $\mathbb{R}$, croissant de 0 à 1 par paliers avec un saut de «hauteur» $\frac{1}{n}$ à chaque fois qu'une valeur observée x_i est atteinte.

On peut également la voir comme la fonction de répartition d'une loi discrète qui donnerait la probabilité $\frac{1}{n}$ à chacune des valeurs $x_1, x_2, ..., x_n$. Cette vision permet de faire le lien entre moment théorique et moment empirique. Le moment théorique peut s'écrire $\mu_r = \int_{\mathbb{R}} x^r dF(x)$ alors que le moment empirique s'écrit en remplaçant F par F_n dans l'intégrale de Riemann-Stieltjes (introduite en note 2.1) : $M_r = \int_{\mathbb{R}} x^r dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} x_i^r$.

5.8 Convergence, approximations gaussiennes, grands échantillons

5.8.1 Les modes de convergence aléatoires

On considère ici une suite infinie de v.a. $\{X_1, X_2, ..., X_n, ...\}$ notée en bref $\{X_n\}$. On peut définir plusieurs modes de convergence pour une telle suite. On notera F_{X_n} la fonction de répartition de X_n.

Définition 5.13 *On dit que $\{X_n\}$* ***converge en loi*** *vers la v.a. X si l'on a, en tout x où sa fonction de répartition F_X est continue,*

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x) ,$$

et l'on note $X_n \xrightarrow[n \to \infty]{\mathcal{L}} X$.

On dira aussi que la loi de X est la *la loi limite* ou *asymptotique* de la suite $\{X_n\}$. En pratique la loi limite sera utile pour donner une **approximation pour le calcul de la probabilité d'un événement sur X_n quand n sera assez grand** :

$$P(X_n \in A) \simeq P(X \in A).$$

Pour la convergence en loi comme pour les autres modes de convergence un cas particulier important est celui où X est une v.a. certaine, c'est-à-dire que la suite converge vers une constante c. Pour la convergence en loi cela implique que $F_{X_n}(x)$ converge vers 0 si $x < c$ et vers 1 si $x \geq c$.

On admettra la proposition suivante, où l'on suppose que les fonctions génératrices existent dans un voisinage de 0.

Proposition 5.13 *La suite de v.a. $\{X_n\}$ converge en loi vers X si et seulement si, pour tout t dans un voisinage de 0, $\lim_{n\to\infty} \Psi_{X_n}(t) = \Psi_X(t)$, où Ψ_{X_n} est la fonction génératrice de X_n et Ψ_X celle de X.*

Cette proposition permet donc d'établir la convergence en loi à partir de la convergence de la fonction génératrice des moments.

Définition 5.14 *On dit que $\{X_n\}$* ***converge en probabilité*** *(ou converge faiblement) vers la v.a. X si, quel que soit $\epsilon > 0$ donné,*

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1\,,$$

et l'on note $X_n \xrightarrow[n\to\infty]{p} X$.

Pour ce mode de convergence comme pour les suivants la convergence vers une constante c s'explicite naturellement en remplaçant X par c.

Définition 5.15 *On dit que $\{X_n\}$* ***converge presque sûrement*** *(ou converge avec probabilité 1, ou converge fortement) vers la v.a. X si, quel que soit $\epsilon > 0$ donné,*

$$\lim_{n\to\infty} P(\sup_{m\geq n}\{|X_m - X|\} < \epsilon) = 1\,,$$

et l'on note $X_n \xrightarrow[n\to\infty]{p.s.} X$.

Cette définition est complexe mais on peut voir qu'elle équivaut à dire que la suite $\{M_n\}$, où $M_n = \sup_{m\geq n}\{|X_m - X|\}$, converge vers 0 en probabilité. Comme pour tout n, $\sup_{m\geq n}\{|X_m - X|\} \geq |X_n - X|$, il est clair que la convergence presque sûre entraîne la convergence en probabilité (d'où les qualificatifs de convergence forte et convergence faible).

On admettra les propositions ci-après, qui pourront nous être utiles par la suite.

Proposition 5.14 *Soit* $\{X_n\}$ *telle que* $X_n \xrightarrow{p.s.} X$ *et* g *une fonction continue alors :*

$$g(X_n) \xrightarrow[n\to\infty]{p.s.} g(X).$$

Proposition 5.15 *Soit* $\{X_n\}$ *telle que* $X_n \xrightarrow{p.s.} X$ *et* $\{Y_n\}$ *telle que* $Y_n \xrightarrow{p.s.} Y$. *Si* f *est continue dans* $\mathbb{R}^2$ *alors :*

$$f(X_n, Y_n) \xrightarrow{p.s.} f(X, Y).$$

Ces deux propositions sont également vraies pour la convergence en probabilité. Elles s'étendent également à des fonctions de k variables aléatoires où $k > 2$.

Définition 5.16 *On dit que* $\{X_n\}$ ***converge en moyenne quadratique*** *vers la v.a.* X *si les v.a.* $X, X_1, X_2, ...$ *ont un moment d'ordre 2 et si*

$$\lim_{n\to\infty} E[(X_n - X)^2] = 0\,,$$

et l'on note $X_n \xrightarrow{m.q.} X$.

La convergence m.q. est particulièrement facile à manipuler car elle repose sur la convergence usuelle d'une suite de nombres $\{E[(X_n - X)^2]\}$. Nous y recourrons abondamment, d'autant plus qu'elle implique la convergence en probabilité.

On admettra la hiérarchie d'implications suivantes (voir certaines démonstrations proposées en exercices) entre les différents modes de convergence :

$$\begin{aligned} p &\Rightarrow \mathcal{L} \\ m.q. &\Rightarrow p \\ p.s. &\Rightarrow p \quad \text{(vu ci-dessus)}. \end{aligned}$$

En outre $p \iff \mathcal{L}$ dans le cas de la convergence vers une constante. Notons que, dans le cas général, il n'y a pas, entre convergence m.q. et convergence p.s., de domination de l'une sur l'autre.

5.8.2 Lois des grands nombres

Théorème 5.3 *Soit* $\{X_n\}$ *une suite de v.a. indépendantes de même loi admettant une moyenne* μ *et une variance* σ^2. *Alors la suite des moyennes empiriques* $\{\bar{X}_n\}$ *converge presque sûrement vers* μ, *i.e. :*

$$\bar{X_n} \xrightarrow[n\to\infty]{p.s.} \mu.$$

Nous énonçons ici la loi dite «forte» des grands nombres. Il existe différentes versions de cette loi requérant des conditions plus ou moins restrictives que celles utilisées ici, dont la loi «faible» concernant la convergence en probabilité. D'un point de vue concret la loi des grands nombres garantit que la moyenne empirique se rapproche de plus en plus de la moyenne de la loi dont est issu l'échantillon quand on augmente la taille de cet échantillon. Aussi, comme on le verra plus loin, la moyenne empirique $\bar{X}_n$ peut-elle prétendre à «estimer» μ.

Historiquement la loi des grands nombres a été introduite par Jakob Bernoulli (publication posthume *Ars conjectandi* en 1713) pour définir la probabilité d'un événement comme étant la limite de sa fréquence relative, au cours d'une série de répétitions d'une expérience aléatoire à l'infini. Il s'agit là du cas particulier où les v.a. $X_1, X_2, ..., X_n, ...$ sont les variables indicatrices de l'occurrence de l'événement (succès) au cours d'un processus de Bernoulli (voir section 4.1.3). Soit $S_n = \sum_{i=1}^{n} X_i$ le nombre total de succès au cours des n premières répétitions, la fréquence relative des occurrences est la moyenne empirique S_n/n et donc :

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{p.s.} p$$

où p est la probabilité de l'événement. La théorie axiomatique moderne des probabilités permet d'établir cette propriété originelle intuitive en précisant des conditions pour qu'elle s'applique, à savoir, dans la version usuelle présentée ici : indépendance des répétitions successives et constance de la probabilité de succès au cours de ces répétitions.

Nous nous bornons à montrer la convergence de $\bar{X}_n$ vers μ en moyenne quadratique qui, rappelons-le, garantit la convergence en probabilité. D'après la proposition 5.1 $E(\bar{X}_n) = \mu$ pour tout n, d'où $E[(\bar{X}_n - \mu)^2] = V(\bar{X}_n) = \frac{\sigma^2}{n}$ qui tend vers 0 quand $n \to \infty$, ce qui établit que $\bar{X}_n \xrightarrow[n\to\infty]{m.q.} \mu$.

La loi des grands nombres n'a pas d'intérêt pratique pour le calcul statistique, contrairement au théorème central limite ci-après qui vient préciser la façon dont $\bar{X}_n$ converge vers μ. Ce théorème est à la base de nombreuses propriétés essentielles des échantillons en statistique.

5.8.3 Le théorème central limite

Théorème 5.4 *Soit $\{X_n\}$ une suite de v.a. indépendantes de même loi admettant une moyenne μ et une variance σ^2. Alors la suite $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converge en loi vers la v.a. de loi $\mathcal{N}(0\,;\,1)$, ce que nous écrivons conventionnellement :*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,;1).$$

Démonstration : nous supposerons que la loi mère admet une fonction génératrice des moments Ψ_X. Une démonstration plus générale considèrerait de même la fonction caractéristique brièvement mentionnée dans la note 2.4 à la fin de la section 2.5, laquelle existe toujours. Dans cette section nous avons également mentionné le développement de $\Psi_X(t)$ en série de Taylor-Mac-Laurin :

$$\Psi_X(t) = \sum_{k=0}^{\infty} \mu_k \frac{t^k}{k!} ,$$

soit, en nous limitant à l'ordre 2,

$$\Psi_X(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2} + o(t^2) = 1 + \mu t + (\sigma^2 + \mu^2)\frac{t^2}{2} + o(t^2)$$

où $\frac{o(t^2)}{t^2} \to 0$ quand $t \to 0$. Soit maintenant :

$$\frac{\bar{X_n} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \sum_{i=1}^n T_i \quad \text{où} \quad T_i = \frac{X_i - \mu}{\sigma\sqrt{n}}.$$

Pour tout i, $X_i - \mu$ a pour moyenne 0 et pour variance σ^2, d'où :

$$\Psi_{X_i-\mu}(t) = 1 + \sigma^2 \frac{t^2}{2} + o(t^2)$$

et
$$\begin{aligned} \Psi_{T_i}(t) &= E\left(e^{(X_i-\mu)\frac{t}{\sigma\sqrt{n}}}\right) \\ &= \Psi_{X_i-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) \\ &= 1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right). \end{aligned}$$

D'après la proposition (3.12),

$$\Psi_{\sum_{i=1}^n T_i}(t) = [1 + \frac{t^2}{2n} + o(\frac{t^2}{n})]^n = (1 + \frac{t^2}{2n})^n + o(\frac{t^2}{n})$$

et, sachant que $\lim_{n\to\infty}(1 + \frac{a}{n})^n = e^a$, on a :

$$\lim_{n\to\infty} \Psi_{\sum_{i=1}^n T_i}(t) = e^{\frac{t^2}{2}} ,$$

qui est bien la fonction génératrice de la loi $\mathcal{N}(0\,;1)$. □

S'il est clair que, pour tout n, la v.a. $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ est centrée-réduite (moyenne nulle, variance égale à 1) le théorème central limite indique en plus que sa loi

tend à être gaussienne quand n s'accroît et ceci, **quelle que soit la loi mère** des X_i.

Application fondamentale

Soit $\bar{X}_n$ la moyenne empirique d'un n-échantillon aléatoire de loi mère quelconque, de moyenne μ et de variance σ^2. Alors si n est assez grand $\bar{X}_n$ suit approximativement une loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ ce que l'on note :

$$\bar{X}_n \underset{approx}{\rightsquigarrow} \mathcal{N}(\mu, \frac{\sigma^2}{n}) .$$

Dans tous les cas (ou presque, voir ci-après pour la loi mère de Bernoulli) $n \geq 30$ suffit pour obtenir des approximations de probabilités à 10^{-2} près. Pour une loi continue à un seul mode sans queues de distribution trop allongées $n = 5$ pourra même suffire. Si la loi mère est gaussienne nous avons vu en section 5.2 que $\bar{X}_n$ est exactement gaussienne pour tout n.

Note 5.5 Comme pour la loi des grands nombres il existe différentes versions du théorème central limite partant de conditions plus ou moins restrictives. En particulier il n'est pas nécessaire que les v.a. soient de même loi ni même qu'elles soient indépendantes dans la mesure où leur degré de dépendance reste faible. Ceci explique que certains phénomènes naturels répondent bien à un modèle gaussien du fait que la variable étudiée résulte de l'addition d'effets aléatoires multiples.

Ainsi on peut établir un comportement asymptotique gaussien pour d'autres types de statistiques dans la mesure où elles sont des moyennes de v.a. qui, sans être nécessairement indépendantes pour n fini, tendent à être i.i.d. quand $n \to \infty$. En particulier ceci est vrai pour la variance de l'échantillon S_n^2 pour laquelle les éléments $X_i - \bar{X}_n$ (et donc leurs carrés) tendent à devenir indépendants du fait que $\bar{X}_n$ converge vers μ. Il est toutefois nécessaire que la variance de S_n^2 existe et il suffit pour cela (voir en fin de section 2.3) que μ_4 existe pour la loi mère (pour le calcul de la variance de la distribution d'échantillonnage de S_n^2 voir les exercices).

Cas particulier : processus de Bernoulli

Soit S_n le nombre total de succès au cours de n répétitions. Comme $E(S_n) = np$ et $V(S_n) = np(1-p)$ on a pour la fréquence relative S_n/n une moyenne p et une variance $\frac{p(1-p)}{n}$. D'où, pour n suffisamment grand,

$$\frac{S_n}{n} \underset{approx}{\rightsquigarrow} \mathcal{N}(p, \frac{p(1-p)}{n})$$

$$\text{ou encore} \quad S_n \underset{approx}{\rightsquigarrow} \mathcal{N}(np, np(1-p)) .$$

Cette deuxième forme constitue **l'approximation de la loi binomiale $\mathcal{B}(n,p)$ par la loi de Gauss** $\mathcal{N}(np, np(1-p))$. C'est l'approche historique qui a permis à de Moivre pour $p = \frac{1}{2}$, puis Laplace pour p quelconque, de mettre initialement en évidence la loi de Gauss (voir exercice 4.3).

En pratique on admet généralement que l'approximation est satisfaisante dès lors que $np \geq 5$ et $n(1-p) \geq 5$. Ces deux conditions garantissent que la moyenne de la loi binomiale ne soit ni trop proche de 0, ni trop proche de n, car dans le cas contraire la loi serait assez nettement asymétrique. Du fait que l'on passe d'une loi discrète à une loi continue on introduit une *correction de continuité* de la façon suivante.

Soit $X \rightsquigarrow \mathcal{B}(n,p)$ alors :

$$P(X = k) \simeq P(k - \frac{1}{2} < U < k + \frac{1}{2}) \text{ où } U \rightsquigarrow \mathcal{N}(np, np(1-p)).$$

Exemple 5.1 Soit $X \rightsquigarrow \mathcal{B}(20\,;0{,}3)$. Nous pouvons recourir à une approximation gaussienne car $np = 6 > 5$ et $n(1-p) = 14 > 5$. Considérons $P(X = 8)$ et $P(X \leq 8)$.

$$\begin{aligned} P(X = 8) &\simeq P(7{,}5 < U < 8{,}5) \text{ où } U \rightsquigarrow \mathcal{N}(6\,;4{,}2) \\ &\simeq P(\frac{7{,}5-6}{\sqrt{4{,}2}} < Z < \frac{8{,}5-6}{\sqrt{4{,}2}}) \text{ où } Z \rightsquigarrow \mathcal{N}(0\,;1) \\ &\simeq P(0{,}73 < Z < 1{,}22) = 0{,}8888 - 0{,}7673 = 0{,}1215. \end{aligned}$$

La valeur exacte (lue dans une table binomiale) est 0,1144 .

$$\begin{aligned} P(X \leq 8) &\simeq P(U < 8{,}5) \\ &\simeq P(Z < 1{,}22) \ = 0{,}8888. \end{aligned}$$

La valeur exacte est 0,8866. ■

Remarque : toutes les lois qui peuvent être définies comme résultant d'une somme de variables aléatoires i.i.d. tendent à être gaussiennes quand le nombre de termes augmente. C'est évidemment le cas de la binomiale $\mathcal{B}(n\ ,p)$ quand $n \to \infty$, comme nous venons de le voir (et par voie de conséquence pour la loi hypergéométrique quand $N \to \infty$ et $M/N \to p$, voir section 4.1.5), mais aussi de la loi binomiale négative $\mathcal{BN}(r,p)$ quand $r \to \infty$, de la loi $\Gamma(r\ ,\lambda)$ quand $r \to \infty$, de la loi $\chi^2(\nu)$ quand $\nu \to \infty$.

De façon indirecte c'est également vrai pour la **loi de Poisson** qui peut être approchée par une somme de v.a. de Bernoulli en découpant l'unité de temps en petits intervalles (voir section 4.1.7). En pratique on peut approcher la loi $\mathcal{P}(\lambda)$ par la loi $\mathcal{N}(\lambda\ ,\lambda)$ dès que $\lambda \geq 20$, les calculs de probabilités étant corrects à 10^{-2} près en utilisant la correction de continuité.

Exemple 5.2 Soit $X \rightsquigarrow \mathcal{P}(20)$. Calculons $P(X \leq 14)$.

$$\begin{aligned} P(X \leq 14) &\simeq P(U < 14{,}5) \text{ où } U \rightsquigarrow \mathcal{N}(20\,;20) \\ &\simeq P(Z < \frac{14{,}5-20}{\sqrt{20}}) \text{ où } Z \rightsquigarrow \mathcal{N}(0;1) \\ &\simeq P(Z < -1{,}23) \ = 0{,}1093. \end{aligned}$$

La valeur exacte (lue dans une table de Poisson) est 0,1049. ■

Nous proposons dans la section des exercices quelques «exercices appliqués» permettant de voir des situations pratiques illustrant l'intérêt des résultats précédents.

5.9 Exercices

Exercice 5.1 Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire de loi mère exponentielle $\mathcal{E}(\lambda)$, montrer que $\bar{X}$ est de loi $\Gamma(n, n\lambda)$.

Exercice 5.2 Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire de loi mère $\Gamma(r\ , \lambda)$. Quelle est la loi de $\bar{X}$?

Exercice 5.3 * Soit $X_1, X_2, ..., X_n$ un échantillon aléatoire de loi mère de Cauchy dont la densité est :

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

Montrer, via la fonction **caractéristique** des moments, que $\bar{X}_n$ suit la même loi (elle ne converge donc pas vers une constante quand $n \to \infty$).

Exercice 5.4 Montrer que la variance de S^2 est égale à $\frac{1}{n}(\mu'_4 - \frac{n-3}{n-1}\sigma^4)$. Que vaut-elle dans le cas particulier de la loi de Gauss (voir formule pour μ'_4 en section 4.2.4) ?

Exercice 5.5 * (Sondage aléatoire simple sans remise) Soit la suite de v.a. $X_1, X_2, \ldots, X_n$ issue du tirage de n individus **sans remise** dans une population de taille N. Soit $a_1, a_2, \ldots, a_N$ les valeurs dans la population de la variable étudiée. Soient $\mu = \frac{1}{N}\sum_{j=1}^{N} a_j$ leur moyenne et $\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(a_j - \mu)^2$ leur variance. Pour des raisons évidentes de symétrie, $P(X_i = a_j)$ reste identique quels que soient i et j.

En déduire la loi marginale de X_i. Par le même type d'argument déterminer la loi d'un couple (X_i, X_k) quels que soient i et k $(i \neq k)$.

Déterminer alors $E(X_i)$, $V(X_i)$ et $E(\overline{X})$ où $\overline{X}$ est la moyenne de l'échantillon sans remise.

Montrer que $cov(X_i, X_k) = -\frac{\sigma^2}{N-1}$ (aide : partir de la formule de décentrage et utiliser la relation générale $(\sum_j a_j)^2 = \sum_j a_j^2 + \sum_{j\neq l} a_j a_l$).

Calculer $V(\sum_{i=1}^{n} X_i)$ (aide : partir de la formule générale donnée à la fin de la section 3.8). En déduire que la moyenne d'un échantillon issu d'un tirage sans remise a pour variance :

$$\frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

(on remarquera que l'on trouve le même facteur correctif de sans remise que pour la loi hypergéométrique en section 4.1.5)

Exercice 5.6 Montrer que la covariance de $\bar{X}$ et de S^2 est égale à $\mu_3^{'}/n$ (aide : on peut supposer que la loi mère est de moyenne nulle sans nuire à la généralité). Ce résultat montre que ces deux statistiques sont asymptotiquement «non corrélées».

Exercice 5.7 Déterminer directement la densité d'une loi $\chi^2(1)$ par le changement de variable de $Z \rightsquigarrow \mathcal{N}(0\,;1)$ à Z^2.

Exercice 5.8 Établir, par la fonction génératrice, la moyenne et la variance de la loi $\chi^2(\nu)$.

Exercice 5.9 * Soit X une v.a. continue de densité f_X, de moyenne μ et de variance σ^2. Soit g une fonction positive.

1. Soit $A = \{x|g(x) > k > 0\}$. Montrer que

$$\int_{-\infty}^{\infty} g(x) f_X(x) dx \geq k \int_A f_X(x) dx.$$

   et en déduire que $E(g(X)) \geq kP(g(X) > k)$.
2. En prenant $g(x) = (x-\mu)^2$ montrer l'*inégalité de Tchebichev :*

$$P(|X-\mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

3. Soit une suite de v.a. $\{Y_n\}$. En prenant $X = |Y_n - Y|$ et $g(x) = x^2$ montrer que $Y_n \xrightarrow{m.q.} Y$ implique $Y_n \xrightarrow{p} Y$.

Exercice 5.10 Démontrer la loi faible des grands nombres quand la variance existe.

Aide : utiliser *l'inégalité de Tchebichev* ci-dessus.

Exercice 5.11 Pour un échantillon de taille n quelle est la probabilité que le maximum dépasse la médiane de la loi mère ? Quelle est la probabilité que le maximum dépasse le troisième quartile (i.e. le quantile d'ordre 0,75) de la loi mère ?

Exercice 5.12 Soit un échantillon de taille n issu d'une loi $\mathcal{U}[0\,;1]$. Déterminer la fonction de répartition et la densité de la loi du minimum de l'échantillon. En déduire l'espérance mathématique de ce minimum.

Exercice 5.13 Pour la marche aléatoire présentée dans la section d'exercices 3.10 donner une valeur approchée, pour n suffisamment grand, de la probabilité d'être éloigné de plus de x mètres de la position initiale après n étapes.

Exercices appliqués

Exercice 5.14 Le niveau de bruit d'un type de machine à laver à un certain régime est une v.a. de moyenne 44 dB et d'écart-type 5 dB. En admettant la validité de l'approximation gaussienne, donner la probabilité de trouver une moyenne supérieure à 48 dB pour un échantillon aléatoire de 10 machines.

Exercice 5.15 Un constructeur automobile indique une consommation de 6,3 l/100km pour un type de modèle dans des conditions expérimentales précises. Pour 30 automobiles (supposées prises au hasard) testées dans ces mêmes conditions on relève une consommation moyenne de 6,42 l/100km et un écart-type de 0,22 l/100km.

Calculer la valeur prise dans cet échantillon par la statistique de Student du théorème 5.2.

A quel quantile correspond-elle sur la loi de cette statistique ? (on supposera que la loi de Student s'applique avec une bonne approximation vu la taille d'échantillon)

L'indication du constructeur vous paraît-elle plausible ?

Exercice 5.16 Un téléphérique a une capacité de 100 personnes. Dans la population française le poids des personnes est distribué avec une moyenne de 66,3 kg et un écart-type de 15,6 kg. En supposant que les personnes entrant dans la benne soient prises au hasard dans cette population quelle est, approximativement, la probabilité que le poids total des personnes transportées excède 7 000 kg ?

Exercice 5.17 Un sondage est effectué auprès de 1 000 personnes dans la population française sur la popularité d'une certaine personnalité.

Quelle est la probabilité que le sondage indique une cote de popularité inférieure ou égale à 42 % si la proportion de personnes favorables à cette personnalité est de 0,44 au sein de la population ? (aide : on aura avantage à passer par la loi de S_n, nombre total de succès, pour pouvoir utiliser la correction de continuité de l'approximation gaussienne, comme dans l'exemple 5.1)

Exercice 5.18 Une machine en fonctionnement normal produit 9 % de pièces défectueuses. Un contrôle de qualité consiste à prélever 120 pièces au hasard. Quelle est la loi du nombre de pièces défectueuses ? Expérience faite, 22 pièces s'avèrent être défectueuses. A quel quantile correspond cette valeur sur la loi précédente ? (aide : on recourra à l'approximation gaussienne avec correction de continuité comme dans l'exemple 5.1)

Qu'en conclure quant au fonctionnement de la machine ?

Exercice 5.19 D'une façon générale on définit la précision d'une méthode de mesure par le double de l'écart-type de son erreur aléatoire. L'hypothèse d'une erreur aléatoire gaussienne est la règle.

Une méthode de mesure d'alcoolémie est réputée avoir une précision de 0,1 mg/l. Sur un même échantillon sanguin on effectue 5 mesures que l'on peut supposer indépendantes. Quelle est la probabilité de trouver un écart-type des 5 mesures supérieur à 0,077 ? (aide : on passera par la variance)

Exercice 5.20 On cherche à prévoir le nombre de nuitées dans les hôtels d'une station balnéaire en juillet. D'expérience on a pu constater que le nombre de nuits passées par un ménage peut être modélisé par une loi de Poisson de moyenne 4. On fait l'hypothèse d'une fréquentation de 10 000 ménages, quelles sont la moyenne et la variance de la v.a. «nombre total de nuitées» ?

En utilisant une approximation gaussienne donner un intervalle de probabilité 0,95 pour cette v.a.

Exercice 5.21 Lors de la conversion du franc à l'euro les opérations sont arrondies au centime d'euro le plus proche. On suppose que les décimales de centime d'euro apparaissent de façon aléatoire uniformément réparties sur l'intervalle $[0\,,1]$. Quelle est approximativement la loi de l'erreur d'arrondi sur 1 000 opérations ? Donner un intervalle de probabilité 0,95 pour cette erreur.

Exercice 5.22 Un appareil électronique contient 3 accumulateurs. Pour que l'appareil fonctionne il faut que les 3 accumulateurs fonctionnent. On admet que la durée de vie d'un accumulateur suit une loi exponentielle de moyenne 2 ans et que les durées des trois éléments sont indépendantes. Quelle est la loi de la durée de fonctionnement de l'appareil ? Quelle est sa moyenne ? Quelle est la probabilité qu'elle soit supérieure à un an ?

Exercice 5.23 Un industriel doit livrer 100 pièces. Sachant que le processus de fabrication produit une pièce défectueuse avec probabilité 0,10 il souhaite budgéter le nombre de pièces à produire pour être quasiment sûr de fournir 100 bonnes unités.

Un raisonnement simpliste consiste à déclarer que 111 pièces suffisent. Quelle est la probabilité de dépasser 111 pièces pour en obtenir 100 bonnes (on pourra utiliser une approximation gaussienne) ?

Combien doit-on fabriquer de pièces pour être sûr d'en avoir 100 bonnes ?
Combien doit-on fabriquer de pièces pour en avoir 100 bonnes avec une probabilité 0,99 ?

Chapitre 6

Théorie de l'estimation paramétrique ponctuelle

6.1 Cadre général de l'estimation

Soit X une v.a. associée à un certain phénomène aléatoire observable de façon répétée comme décrit en section 5.1. Notre objectif est «d'estimer» certaines *caractéristiques* d'intérêt de sa loi (la moyenne, la variance, la fonction de répartition, la fonction de densité, etc.) sur la base d'une série d'observations $x_1, x_2, ..., x_n$. Un cas particulier important est celui du sondage dans une population (voir section 3.7) dont l'objectif est d'estimer diverses caractéristiques descriptives de celle-ci.

Dans ce chapitre nous considérerons toujours, même si des développements analogues sont possibles dans d'autres circonstances, que $x_1, x_2, ..., x_n$ sont des réalisations d'un n-échantillon aléatoire $X_1, X_2, ..., X_n$. Cette hypothèse sur nos observations qui peut être plus ou moins réaliste est nécessaire pour étudier de façon simple, en termes probabilistes, la qualité des estimations que l'on cherche à produire. Ce chapitre ne traite également que du problème de l'*estimation ponctuelle*, c'est-à-dire celle qui consiste à attribuer, au mieux de notre savoir, une valeur unique à la caractéristique d'intérêt inconnue. Au chapitre 7 nous aborderons l'*estimation par intervalle* consistant à donner un encadrement plausible pour la caractéristique.

La théorie de l'estimation étudie les propriétés des estimations et des méthodes générales d'estimation comme celle dite du «maximum de vraisemblance». L'objectif est de comparer les lois d'échantillonnage des «estimateurs» pour établir des préférences lorsque plusieurs choix se présentent. La notion d'estimateur est la notion centrale de ce chapitre alors même qu'elle ne se définit pas formellement en termes mathématiques.

Définition informelle d'un estimateur et d'une estimation

Dans le cadre défini ci-dessus, soit à estimer une caractéristique c de la variable aléatoire X sur la base de la réalisation $(x_1, x_2, ..., x_n)$ du n-échantillon $(X_1, X_2, ..., X_n)$. On appellera *estimateur* toute statistique (donc toute fonction de $X_1, X_2, ..., X_n$, voir définition 5.2) dont la réalisation après expérience est envisagée comme estimation de c. Un estimateur se définit donc dans l'intention de fournir une *estimation*.

Insistons sur le fait qu'un estimateur est une variable aléatoire, alors qu'une estimation est une valeur numérique prise par l'estimateur suite à la réalisation du n-échantillon. Si un estimateur est déterminé par une fonction $h(X_1, X_2, ..., X_n)$, l'estimation correspondante est évidemment $h(x_1, x_2, ..., x_n)$. Soit, par exemple, à estimer la moyenne $E(X)$ de la loi de X, un estimateur qui semble a priori naturel est la moyenne empirique $\bar{X}$ qui produit une estimation $\bar{x}$, moyenne descriptive de la série des valeurs observées.

6.2 Cadre de l'estimation paramétrique

En estimation paramétrique la loi de X est réputée appartenir à une famille de lois, telles que celles présentées au chapitre 4, qui peut être décrite par une forme fonctionnelle connue soit de sa fonction de répartition, soit de sa fonction de densité dans le cas continu, soit de sa fonction de probabilité dans le cas discret, forme dépendant d'un ou plusieurs *paramètres inconnus* réels. De façon générique nous noterons θ ce paramètre ou vecteur de paramètres et $F(x;\theta)$, $f(x;\theta)$ ou $p(x;\theta)$ les trois formes fonctionnelles précitées. Toutefois, par simplification et sauf mention expresse contraire, **nous noterons $f(x;\theta)$ aussi bien la densité du cas continu que la fonction de probabilité du cas discret**.

L'ensemble des valeurs possibles pour θ, appelé *espace paramétrique*, sera noté Θ, lequel est inclus dans $\mathbb{R}^k$ où k est la *dimension* du paramètre θ. Le plus souvent la famille paramétrique à laquelle la loi de X est réputée appartenir sera décrite par la famille de densités de probabilité (respectivement de fonctions de probabilité) $\{f(x;\theta); \theta \in \Theta\}$. Ces fonctions sont implicitement définies pour tout $x \in \mathbb{R}$. Rappelons ici (voir section 1.4) qu'on appelle *support* de $f(x;\theta)$ (ou support de la loi) l'ensemble des valeurs de x telles que $f(x;\theta) > 0$. En termes courants, on parlera de l'ensemble des réalisations (ou valeurs) possibles.

Lorsque nous considérerons une famille de lois usuelles nous reviendrons aux notations du chapitre 4. Ainsi la famille des lois de Gauss est décrite par la famille des densités de la forme $(1/(\sqrt{2\pi}\sigma))\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$, pour tout $x \in \mathbb{R}$, où intervient un paramètre (μ, σ^2) de dimension 2, l'espace paramétrique étant la partie de $\mathbb{R}^2$: $\{(\mu, \sigma^2) \,|\, \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R},\ \sigma^2 > 0\}$.

Dans ce cadre paramétrique le problème est celui de l'estimation du paramètre θ grâce à laquelle on obtiendra une estimation complète de la loi

de X et, par voie de conséquence, de toute caractéristique de cette loi. Distinguons bien ici la notion de paramètre d'une loi de celle de caractéristique (moyenne, variance, médiane, ...) de la loi : le paramètre identifie chaque loi (chaque membre) dans la famille considérée mais n'est pas nécessairement une caractéristique usuelle de cette loi. Par contre toute caractéristique usuelle dépend du membre de la famille et donc du paramètre θ. Aussi le moment d'ordre k, par exemple, sera-t-il noté $\mu_k(\theta)$, la moyenne sera notée $\mu(\theta)$ et la variance $\sigma^2(\theta)$. Si notre objectif principal est d'estimer le paramètre inconnu θ, il se pourra aussi que nous souhaitions directement **estimer une fonction de** θ représentant une certaine caractéristique particulièrement intéressante, sans nécessairement passer par l'estimation de θ. En particulier, fréquemment on voudra estimer moyenne et variance de la loi, soit $\mu(\theta)$ et $\sigma^2(\theta)$.

Notation pour les estimateurs et estimations

Pour un paramètre désigné par une certaine lettre on note souvent un estimateur par la même lettre surmontée d'un accent circonflexe. Pour distinguer la méthode d'estimation utilisée on pourra ajouter en indice supérieur une lettre y faisant clairement référence. Ainsi, pour le paramètre générique θ un estimateur non précisé sera noté $\widehat{\theta}$, l'estimateur obtenu par la méthode des moments (exposée en section 6.4) sera noté $\widehat{\theta}^M$ et l'estimateur obtenu par la méthode du maximum de vraisemblance (exposée en section 6.7) sera noté $\widehat{\theta}^{MV}$. Selon nos conventions initiales nous devrions noter ces variables aléatoires avec la lettre majuscule de θ. Mais dans la mesure où le contexte indique clairement s'il s'agit d'une variable aléatoire ou de sa réalisation, **nous ne ferons pas la distinction entre estimation et estimateur lorsqu'ils sont notés en lettres grecques**. La lettre $\hat{\lambda}^M$ désignera, par exemple, l'estimateur ou l'estimation des moments pour le paramètre λ de la loi $\mathcal{E}(\lambda)$.

Notons que dans le contexte de l'estimation paramétrique le paramètre θ est totalement inconnu. Ainsi la v.a. X : «intervalle de temps séparant r occurrences» dans un processus de Poisson suit une loi $\Gamma(r, \lambda)$ dont seul le paramètre λ est inconnu. Nous avons donc affaire à une sous-famille de la famille des lois Gamma.

Remarquons également qu'il n'y a pas qu'une seule façon de paramétrer une famille de lois. En particulier toute fonction strictement monotone $h(\theta)$ du paramètre θ permet une *reparamétrisation* de la famille des densités (respectivement des fonctions de probabilité). Ainsi, nous avons adopté, pour décrire la famille des lois exponentielles $\mathcal{E}(\lambda)$, la forme fonctionnelle :

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

dans laquelle λ est **l'inverse** de la moyenne de la loi mais correspond à l'intensité du processus de Poisson (nombre moyen d'occurrences par unité de temps).

Certains auteurs utilisent la forme :

$$f(x;\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}, x \geq 0, \theta > 0$$

où $\theta = 1/\lambda$ est la fonction de reparamétrisation, auquel cas θ est la moyenne de la loi. Nous garderons à l'esprit ce problème de changement de paramètre qui permettra de dégager certaines propriétés intéressantes des estimateurs.

Pour clore cette introduction mettons en évidence le fait que, dans le cadre des échantillons aléatoires, la loi conjointe du n-échantillon $(X_1, X_2, ..., X_n)$ peut être définie par la fonction de densité (respectivement la fonction de probabilité) conjointe :

$$f(x_1, x_2, .., ., x_n; \theta) = f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

où θ est le paramètre inconnu dans Θ (par commodité, nous désignons ici, et ferons de même éventuellement plus loin, la densité conjointe par la même lettre f que la densité de la loi mère). Dans le cas discret cette expression est la probabilité $P_\theta(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$.

Pour établir certains résultats nous serons amenés à poser des conditions sur cette densité (ou fonction de probabilité) conjointe. Ces conditions ne seront pertinentes qu'aux points $(x_1, x_2, .., ., x_n)$ de $\mathbb{R}^n$ correspondant à des valeurs possibles, en d'autres termes uniquement sur le support de la loi conjointe. Ce support est évidemment l'ensemble produit n fois du support de $f(x;\theta)$. Pour la famille paramétrique dans son ensemble, il faudra prendre en compte l'union des supports de tous les membres lorsque θ décrit Θ. Pour la famille des lois $\mathcal{U}[0, \theta]$, avec $\theta > 0$, par exemple, cette union est $\mathbb{R}^+$. Donc, par la suite, les conditions imposées aux densités ou fonctions de probabilités seront implicitement restreintes à leurs supports.

6.3 La classe exponentielle de lois

Cette classe regroupe des **familles paramétriques de lois** qui, de par leur forme particulière, partagent beaucoup de propriétés dans la théorie de l'estimation ou la théorie des tests, du fait que leurs densités peuvent s'écrire sous une même expression canonique (on parle aussi de la «famille exponentielle» mais cela prête à confusion avec la famille exponentielle usuelle $\mathcal{E}(\lambda)$).

Définition 6.1 *Soit une famille paramétrique de lois admettant des fonctions de densité (cas continu) ou des fonctions de probabilité (cas discret)* $\{f(x;\theta); \theta \in \Theta \subseteq \mathbb{R}^k\}$. *On dit qu'elle appartient à la* ***classe exponentielle*** *de lois si* $f(x;\theta)$ *peut s'écrire :*

$$f(x;\theta) = a(\theta) b(x) \exp\{c_1(\theta) d_1(x) + c_2(\theta) d_2(x) + ... + c_k(\theta) d_k(x)\}$$

pour tout $x \in \mathbb{R}$.

Notons qu'il doit y avoir autant de termes de produits dans la partie exponentielle que de dimensions pour θ. En particulier, si θ est de dimension 1, on a :

$$f(x;\theta) = a(\theta)b(x)\exp\{c(\theta)d(x)\}.$$

De plus cette forme canonique doit être effective **pour tout** $x \in \mathbb{R}$. En particulier si le support de $f(x;\theta)$, donc l'ensemble des valeurs de x pour lesquelles $f(x;\theta) > 0$, dépend lui-même du paramètre θ, un terme d'indicatrice $I_{[\alpha(\theta),\beta(\theta)]}(x)$ doit être introduit, ce qui ne peut en aucun cas permettre la forme définie ci-dessus. Ces types de lois auront, de ce fait, des propriétés très spécifiques. On donnera pour exemple la famille des lois $\mathcal{U}[0,\theta]$, où $\theta > 0$ est inconnu, dont les densités sont de la forme :

$$f(x;\theta) = \frac{1}{\theta} I_{[0,\theta]}(x) \ , \forall x \in \mathbb{R}.$$

Notons encore que les notions de dimension k et de forme canonique ne concernent que les paramètres inconnus. Ainsi la sous-famille des lois $\mathcal{B}(n,p)$ où n est connu, comme c'est toujours le cas dans les applications statistiques, appartient à la famille exponentielle (voir exemple 6.1), ce qui ne serait pas le cas si n était inconnu.

Nous verrons plus loin que les fonctions $d_i(x)$ jouent un rôle central dans la recherche des meilleurs estimateurs. Aussi les mettons-nous en évidence dans les trois exemples qui suivent, puis dans le tableau 6.1 qui servira de référence par la suite.

Exemple 6.1 Loi $\mathcal{B}(n,p)$ avec n connu.

$$\begin{aligned} f(x;n,p) &= \binom{n}{x} p^x (1-p)^{n-x} \quad \text{pour} \quad x = 0,1,2,...,n \\ &= (1-p)^n \binom{n}{x} \exp\{x.\ln\frac{p}{1-p}\} \end{aligned}$$

d'où $d(x) = x$. Le cas de la loi de Bernoulli $\mathcal{B}(p)$ est identique avec $n = 1$. ■

Exemple 6.2 Loi $\mathcal{P}$ (λ).

$$\begin{aligned} f(x;\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{pour} \quad x \in \mathbb{N} \\ &= e^{-\lambda}\frac{1}{x!}\exp\{x.\ln\lambda\} \end{aligned}$$

d'où, également, $d(x) = x$. ■

Exemple 6.3 Loi $\mathcal{N}(\mu\ ,\sigma^2)$.

$$f(x;\mu\ ,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\},\ \ x\in\mathbb{R}$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{\mu^2}{2\sigma^2}\}\exp\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x\}$$

d'où $d_1(x)=x^2$ et $d_2(x)=x$. ■

On peut établir aisément les résultats du tableau suivant qui contient la plupart des lois usuelles.

Tableau 6.1 : Principales lois usuelles appartenant à la classe exponentielle

loi	paramètre	$d_1(x)$	$d_2(x)$	I [1]
$\mathcal{B}(p)$	p	x	-	$[p(1-p)]^{-1}$
$\mathcal{B}(n,p)$	$p\,(n$ connu$)$	x	-	$n[p(1-p)]^{-1}$
$\mathcal{BN}(r,p)$	$p\,(r$ connu$)$	x	-	$r[p^2(1-p)]^{-1}$
$\mathcal{P}(\lambda)$	λ	x	-	$1/\lambda$
$\mathcal{E}(\lambda)$	λ	x	-	$1/\lambda^2$
$\Gamma(r,\lambda)$	$\lambda\,(r$ connu$)$	x	-	r/λ^2
$\mathcal{N}(\mu,\sigma^2)$	(μ,σ^2)	x^2	x	
$Pareto(a,\theta)$	$\theta\,(a$ connu$)$	$\ln(x)$	-	$1/\theta^2$
$Beta(\alpha,\beta)$	(α,β)	$\ln(x)$	$\ln(1-x)$	

[1] *Information de Fisher, voir section 6.6.3*

En revanche les lois hypergéométriques (M inconnu), Weibull et Gumbel n'appartiennent pas à la classe exponentielle.

6.4 Une approche intuitive de l'estimation : la méthode des moments

Bien que cette méthode ne soit pas toujours satisfaisante nous l'introduisons dès maintenant en raison de son côté intuitif. Elle nous servira ainsi, dans la section suivante, à illustrer les propriétés générales des estimateurs.

Nous commençons par le cas d'un paramètre à une dimension. Pour une réalisation $x_1, x_2, ..., x_n$ de l'échantillon la méthode consiste alors à choisir pour estimation de θ la valeur telle que la moyenne théorique $\mu(\theta)$ (ou premier moment de la loi) coïncide avec la moyenne empirique $\overline{x}$. Pour la loi $\mathcal{E}(\lambda)$, par exemple, l'estimation de λ sera $\hat{\lambda}^M$ telle que $1/\hat{\lambda}^M = \bar{x}$, soit $\hat{\lambda}^M = 1/\bar{x}$. Pour la loi $\mathcal{BN}(r,p)$ avec r connu, l'estimation de p sera $\hat{p}^M$ telle que

$$\frac{r(1-\hat{p}^M)}{\hat{p}^M} = \bar{x},\ \text{d'où}\ \ \hat{p}^M = \frac{r}{r+\bar{x}}.$$

Pour la loi de Poisson la solution est $\bar{x}$ puisque le paramètre λ est lui-même la moyenne de la loi. De même, pour la loi de Bernoulli, p est estimé par la moyenne qui est la fréquence relative observée.

La méthode n'a de sens que s'il y a existence et unicité de la solution dans l'espace paramétrique Θ, ce que nous supposerons toujours vrai. Ainsi, de façon générale, nous sommes amenés à résoudre l'équation $\mu(\theta) = \bar{x}$ et, en raison de l'unicité, nous pourrons noter $\hat{\theta}^M = \mu^{-1}(\bar{x})$ l'estimation de θ pour une réalisation $\bar{x}$ donnée de $\bar{X}$. Appliquée maintenant à $\bar{X}$ (donc dans l'univers - selon l'acceptation de ce mot donnée en fin de section 5.1 - des échantillons aléatoires) la fonction μ^{-1} définit alors la statistique $\hat{\theta}^M = \mu^{-1}(\bar{X})$ appelée *estimateur des moments* de θ (rappelons que nous n'utilisons pas de lettre grecque majuscule pour distinguer estimateur et estimation).

Pour un paramètre de dimension 2 l'estimation résulte de la résolution de deux équations, l'une reposant sur le premier moment, l'autre sur le moment d'ordre 2. Prenons le cas de la loi de Gauss avec (μ, σ^2) comme paramètre, dont le premier moment est μ lui-même et le moment d'ordre 2 est $E(X^2) = \mu^2 + \sigma^2$. On résout donc (en passant directement aux v.a.)

$$\begin{cases} \mu & = \bar{X} \\ \mu^2 + \sigma^2 & = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \end{cases}$$

d'où $\hat{\mu}^M = \bar{X}$ et $\hat{\sigma}^{2,M} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \widetilde{S}^2$. La moyenne et la variance théoriques sont donc estimées naturellement par la moyenne et la variance empiriques.

Prenons maintenant le cas moins intuitif de la loi de Gumbel de paramètre (α, β), dont la moyenne est $\alpha + \gamma\beta$, où γ est la constante d'Euler, et la variance est $\pi^2\beta^2/6$ (voir section 4.2.8). On résout :

$$\begin{cases} \alpha + \gamma\beta & = \bar{X} \\ (\alpha + \gamma\beta)^2 + \dfrac{\pi^2\beta^2}{6} & = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \end{cases}$$

ou, de façon équivalente,

$$\begin{cases} \alpha + \gamma\beta & = \bar{X} \\ \dfrac{\pi^2\beta^2}{6} & = \widetilde{S}^2 \end{cases}$$

ce qui donne la solution $\hat{\beta}^M = \frac{\sqrt{6}}{\pi} \widetilde{S}$ et $\hat{\alpha}^M = \bar{X} - \frac{\gamma\sqrt{6}}{\pi} \widetilde{S}$.

D'une façon générale l'estimation de θ de dimension 2 par la méthode des moments est la solution (supposée exister et être unique pour toute réalisation du n-échantillon aléatoire) du système :

$$\begin{cases} \mu(\theta) & = \bar{x} \\ \mu_2(\theta) & = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{cases} .$$

Cette solution appliquée à $\bar{X}$ et $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ donne l'estimateur de θ correspondant.

Du fait de la correspondance des formules de décentrage pour les moments empiriques et pour les moments théoriques il est équivalent de résoudre :

$$\begin{cases} \mu(\theta) & = \bar{x} \\ \sigma^2(\theta) & = \widetilde{s}^2 \end{cases}$$

où la deuxième équation porte donc sur les moments centrés d'ordre 2. Nous donnons maintenant une définition formelle de l'estimateur des moments dans le cas général où le paramètre est de dimension k quelconque.

Définition 6.2 *Soit un échantillon aléatoire $(X_1, X_2, ..., X_n)$ dont la loi mère appartient à une famille paramétrique de paramètre inconnu $\theta \in \Theta$, où $\Theta \subseteq \mathbb{R}^k$, et telle que pour tout $\theta \in \Theta$ il existe un moment $\mu_k(\theta)$ à l'ordre k. Si, pour toute réalisation $(x_1, x_2, ..., x_n)$ de $(X_1, X_2, ..., X_n)$ le système à k équations*

$$\begin{cases} \mu_1(\theta) & = m_1 \\ \mu_2(\theta) & = m_2 \\ & ... \\ \mu_k(\theta) & = m_k \end{cases}$$

(où m_r dénote la réalisation du moment empirique d'ordre r : $m_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$) admet une solution unique, cette solution est appelée estimation des moments de θ. La fonction(de $\mathbb{R}^n$ dans $\mathbb{R}^k$) qui à toute réalisation $(x_1, x_2, ..., x_n)$ fait correspondre cette solution définit, en s'appliquant à $(X_1, X_2, ..., X_n)$, une statistique à valeurs dans $\mathbb{R}^k$ appelée ***estimateur des moments*** *de θ.*

6.5 Qualités des estimateurs

Un des objectifs essentiels de la théorie de l'estimation, nous l'avons dit, est d'opérer des choix parmi les différents estimateurs auxquels on peut penser. Pour cela il est nécessaire de se donner des critères de qualité pertinents. De façon générique nous noterons T_n l'estimateur de θ à étudier. Étant donné que la valeur de θ est inconnue, nous souhaitons que le comportement de T_n soit satisfaisant quel que soit $\theta \in \Theta$, c'est-à-dire quelle que soit la loi mère effective dans la famille paramétrique donnée, et les critères de qualité seront à étudier comme des **fonctions de** θ. Les critères définis ci-après, mis à part

l'exhaustivité (section 6.5.4), seront appliqués uniquement à un paramètre de dimension 1 ($\Theta \subseteq \mathbb{R}$) et nous commenterons en section 6.6.4 les possibilités d'extension à une dimension supérieure.

6.5.1 Biais d'un estimateur

Définition 6.3 *Soit une v.a. X de loi de densité (ou fonction de probabilité) $f(x;\theta)$ où $\theta \in \Theta \subseteq \mathbb{R}$. Soit $X_1, X_2..., X_n$ un n-échantillon issu de cette loi et T_n un estimateur de θ. On appelle* ***biais*** *de T_n pour θ la valeur :*

$$b_\theta(T_n) = E_\theta(T_n) - \theta.$$

Si $b_\theta(T_n) = 0$ quel que soit $\theta \in \Theta$, on dit que T_n est ***sans biais*** *pour θ.*

Cette définition s'étend naturellement à l'estimation d'une fonction $h(\theta)$. Le biais caractérise donc l'écart entre la moyenne de T_n dans l'univers de tous les échantillons possibles et la valeur cible θ. Elle correspond à la notion d'erreur systématique pour un instrument de mesure. Notons que la moyenne de T_n, $E_\theta(T_n)$, est indicée par θ pour rappeler qu'elle est liée à la valeur inconnue de θ. Ceci sera vrai également plus loin pour la variance et l'erreur quadratique moyenne de T_n. Pour alléger les écritures, nous omettrons pourtant souvent cette indexation, notamment dans les illustrations.

Exemple 6.4 Soit la famille des lois continues $U[0,\theta]$. Montrons que $\frac{n+1}{n}X_{(n)}$ est sans biais pour θ. Nous avons vu en section 5.6 que, pour une loi de fonction de répartition $F(x)$, la fonction de répartition du maximum de l'échantillon $X_{(n)}$ est $[F(x)]^n$. Dans la situation particulière considérée $F(x;\theta) = \frac{x}{\theta}$ quand $x \in [0,\theta]$ et la densité de $X_{(n)}$ est donc :

$$\frac{1}{\theta}.n\left(\frac{x}{\theta}\right)^{n-1} = n\frac{x^{n-1}}{\theta^n} \quad \text{si } x \in [0,\theta],$$

d'où :

$$E_\theta(X_{(n)}) = \int_0^\theta x.n\frac{x^{n-1}}{\theta^n}dx = \frac{n}{n+1}\frac{\theta^{n+1}}{\theta^n} = \frac{n}{n+1}\theta.$$

Donc $E_\theta(\frac{n+1}{n}X_{(n)}) = \theta$ quel que soit θ. Il est intéressant de noter la présence du facteur $\frac{n+1}{n}$ qui, en quelque sorte, prend en compte l'écart entre le maximum observé et la borne supérieure des valeurs possibles. ■

Exemple 6.5 Considérons l'estimation des moments de la loi mère qui sont des **fonctions de** θ. Remarquons que le moment empirique d'ordre r est sans biais pour le moment théorique d'ordre r de la loi de X. En effet, par définition, $E_\theta(X^r) = \mu_r(\theta)$ et donc, quel que soit θ,

$$E_\theta(\frac{1}{n}\sum_{i=1}^n X_i^r) = \frac{1}{n}\sum_{i=1}^n E_\theta(X_i^r) = \frac{1}{n}n\mu_r(\theta) = \mu_r(\theta).$$

Ceci n'est pas vrai pour les moments centrés. Ainsi pour $r = 2$ nous avons la variance empirique $\widetilde{S}^2$ pour laquelle (voir proposition 5.2) $E_\theta(\widetilde{S}^2) = \frac{n-1}{n}\sigma^2(\theta)$. Son biais est :

$$E_\theta(\widetilde{S}^2) = \frac{n-1}{n}\sigma^2(\theta) - \sigma^2(\theta) = -\frac{1}{n}\sigma^2(\theta).$$

Ceci signifie, qu'en moyenne, la variance empirique sous-estime la variance de la loi étudiée. C'est pourquoi on lui préfère la statistique

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

appelée conventionnellement «variance de l'échantillon» qui est sans biais pour θ (voir section 5.2). Rappelons que cette sous-estimation s'explique par le fait que les écarts sont mesurés par rapport à la moyenne même des valeurs et non par rapport à la vraie moyenne $\mu(\theta)$. ■

Il est intéressant de remarquer que ces propriétés des moments sont vraies pour toute loi mère (dans la mesure où les moments existent) indépendamment de tout cadre paramétrique. On dit que ce sont des propriétés *non paramétriques* (en anglais : *distribution free*) que nous développerons au chapitre 8.

6.5.2 Variance et erreur quadratique moyenne d'un estimateur

La variance $V_\theta(T_n)$ de l'estimateur est un critère important dans la mesure où elle caractérise la dispersion des valeurs de T_n dans l'univers des échantillons possibles. Toutefois il s'agit de la dispersion autour de $E_\theta(T_n)$ et non pas autour de θ. Pour prendre en compte l'écart par rapport à θ on introduit le critère d'erreur quadratique moyenne.

Définition 6.4 *On appelle* ***erreur quadratique moyenne*** *de T_n par rapport à θ, la valeur, notée $eqm_\theta(T_n)$, définie par :*

$$eqm_\theta(T_n) = E_\theta[(T_n - \theta)^2],$$

et l'on a :

$$eqm_\theta(T_n) = [b_\theta(T_n)]^2 + V_\theta(T_n).$$

En effet :

$$\begin{aligned} E_\theta[(T_n - \theta)^2] &= E_\theta[\{T_n - E_\theta(T_n) + E_\theta(T_n) - \theta\}^2] \\ &= E_\theta[\{T_n - E_\theta(T_n)\}^2] + [E_\theta(T_n) - \theta]^2 + 2E_\theta[T_n - E_\theta(T_n)][E_\theta(T_n) - \theta] \\ &= V_\theta(T_n) + [b_\theta(T_n)]^2 \quad \text{car } E_\theta[T_n - E_\theta(T_n)] = 0. \end{aligned}$$

Comme l'indique son nom ce critère mesure la **distance au carré** à laquelle T_n se situe en moyenne par rapport à θ. On peut faire l'analogie avec les impacts effectués par un tireur sur une cible (même si cela correspond plutôt à un paramètre de dimension 2). Le tireur cherche à atteindre le centre de la cible mais ses impacts, au cours des répétitions («univers» de ses tirs), peuvent être systématiquement décalés, c'est-à-dire que le centre de ceux-ci n'est pas le centre de la cible. En revanche ses tirs peuvent être très groupés (variance faible). Un autre tireur peut être bien centré (biais nul ou faible) mais avoir peu de régularité et donc une forte dispersion de ses tirs (variance élevée). Le choix du meilleur tireur dépend de l'importance relative du décalage systématique et de la régularité.

Le critère d'erreur quadratique moyenne (en bref e.q.m.) n'est pas la panacée mais il est préféré parce qu'il s'exprime en fonction des notions simples de biais et de variance. D'autres critères peuvent paraître tout aussi naturels, en particulier l'erreur absolue moyenne $E_\theta(|T_n - \theta|)$, mais celle-ci est beaucoup plus difficile à manipuler analytiquement.

En adoptant le critère d'e.q.m. pour juger de la précision d'un estimateur le problème est de rechercher le meilleur estimateur au sens de ce critère, ce qui nous conduit aux définitions suivantes.

Définition 6.5 *On dit que l'estimateur T_n^1* ***domine*** *l'estimateur T_n^2 si pour tout $\theta \in \Theta$, $eqm_\theta(T_n^1) \leq eqm_\theta(T_n^2)$, l'inégalité étant stricte pour au moins une valeur de θ.*

L'idéal serait de disposer d'un estimateur qui domine tous les autres. Or il n'existe pas en général, d'estimateur d'e.q.m. minimale *uniformément* en θ. Pour s'en convaincre considérons comme estimateur la v.a. certaine θ_0 où θ_0 est l'une des valeurs possibles. Pour celui-ci l'e.q.m. en $\theta = \theta_0$ est nulle alors que pour tout autre estimateur l'e.q.m. est strictement positive (au moins par sa variance s'il est véritablement aléatoire ou par son biais s'il est certain). Cet estimateur particulier ne peut donc être dominé. Néanmoins, si un estimateur est dominé par un autre estimateur, il n'est pas utile de le retenir.

Définition 6.6 *On dit qu'un estimateur est* ***admissible*** *s'il n'existe aucun estimateur le dominant.*

Ainsi seule est à prendre en compte la classe des estimateurs admissibles. A partir de là, plusieurs orientations de choix sont possibles, l'une des plus répandues étant de choisir l'estimateur pour lequel le maximum que peut atteindre l'e.q.m. sur Θ est le plus faible.

Définition 6.7 *On dit que T_n^* est* ***minimax*** *si pour tout autre estimateur T_n on a :*

$$\sup_{\theta \in \Theta} eqm_\theta(T_n^*) \leq \sup_{\theta \in \Theta} eqm_\theta(T_n).$$

Nous ne poursuivons pas ici la recherche d'estimateurs minimax et nous nous contenterons d'illustrer par deux exemples les propriétés de dominance et d'admissibilité.

Exemple 6.6 Soit une loi mère $\mathcal{N}(\mu, \sigma^2)$ et un échantillon de taille $n > 1$. Montrons que, pour estimer σ^2, $\widetilde{S}^2 = \frac{(n-1)}{n} S^2$ domine S^2. Pour ce dernier le biais est nul et la variance est $\frac{2\sigma^4}{n-1}$ (voir section 5.3). D'où :

$$\begin{aligned} E(\widetilde{S}^2) &= E\left(\frac{(n-1)S^2}{n}\right) = \frac{n-1}{n}\sigma^2 \\ V\left(\frac{(n-1)S^2}{n}\right) &= \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2} \\ eqm\left(\frac{(n-1)S^2}{n}\right) &= \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 + \frac{2(n-1)\sigma^4}{n^2} \\ &= \frac{1}{n^2}\sigma^4 + \frac{2(n-1)\sigma^4}{n^2} = \frac{(2n-1)\sigma^4}{n^2}\,. \end{aligned}$$

La différence $eqm(S^2) - eqm(\widetilde{S}^2)$ est donc :

$$\frac{2\sigma^4}{n-1} - \frac{(2n-1)\sigma^4}{n^2} = \frac{(3n-1)\sigma^4}{(n-1)n^2}$$

qui est toujours positif. Par conséquent S^2 n'est pas admissible.
En fait $\widetilde{S}^2$ introduit un biais, mais celui-ci (au carré) est compensé par une variance plus faible. Notons que ceci n'est pas vrai pour toute loi mère (voir exercices). ■

Exemple 6.7 Soit à estimer le paramètre p d'une loi de Bernoulli (ou, en situation pratique, une proportion p par sondage dans une population). Soit $S_n = \sum_{i=1}^n X_i$ le total empirique ou fréquence de succès observée. Montrons que si p est au voisinage de $1/2$ la statistique $T = (S_n+1)/(n+2)$ est préférable, au sens de l'e.q.m., à la proportion empirique naturelle S_n/n pour estimer p. Comme S_n suit une loi $B(n,p)$, on a $E(S_n) = np$ et $V(S_n) = np(1-p)$. Pour la proportion empirique $E(S_n/n) = p$, le biais est donc nul et l'e.q.m. est égale à $\frac{p(1-p)}{n}$. Pour le deuxième estimateur T, on a :

$$E(T) = \frac{np+1}{n+2} \quad \text{et} \quad V(T) = \frac{V(S_n)}{(n+2)^2} = \frac{np(1-p)}{(n+2)^2}$$

d'où son e.q.m. :

$$eqm(T) = \left[\frac{np+1}{n+2} - p\right]^2 + \frac{np(1-p)}{(n+2)^2} = \frac{(1-2p)^2 + np(1-p)}{(n+2)^2}.$$

En faisant le rapport de cette e.q.m. à celle de S_n/n on obtient :

$$\frac{n}{(n+2)^2}\left[n+\frac{(1-2p)^2}{p(1-p)}\right].$$

Or pour $p=\frac{1}{2}$ ceci vaut $\frac{n^2}{(n+2)^2}<1$ et le rapport ci-dessus étant une fonction continue de p dans $]0,1[$, il reste strictement inférieur à 1 dans un certain voisinage de 1/2. Un calcul plus approfondi montrerait que ce voisinage dépend de n et est l'intervalle

$$\left]\frac{1}{2}-\sqrt{\frac{n+1}{2n+1}}\,,\,\frac{1}{2}+\sqrt{\frac{n+1}{2n+1}}\right[.$$

En conclusion, aucun des deux estimateurs ne domine l'autre. ■

Dans ces deux exemples on constate que si l'on accepte un certain biais, des estimateurs apparemment naturels peuvent être moins performants au sens de l'e.q.m.. Toutefois de nombreux statisticiens privilégient les estimateurs sans biais signifiant ainsi qu'ils ne considèrent pas l'e.q.m. comme la panacée. Si l'on se restreint à la classe des estimateurs sans biais des résultats tangibles peuvent être obtenus dans la recherche de l'estimateur optimal et ceux-ci seront présentés en section 6.6.

6.5.3 Convergence d'un estimateur

Nous considérons ici la suite $\{T_n\}$ de v.a. à valeurs dans $\mathbb{R}$ lorsque la taille n de l'échantillon s'accroît à l'infini, toujours avec $\Theta\subseteq\mathbb{R}$. Pour un estimateur digne de ce nom on s'attend à ce qu'il se rapproche de plus en plus de θ quand $n\to\infty$. C'est ce qu'exprime la notion de convergence. Formellement on dira que l'estimateur T_n est *convergent* selon un certain mode «m» si :

$$T_n \xrightarrow[n\to\infty]{\text{«m»}} \theta$$

où «m» est à remplacer par p, p.s ou m.q. respectivement pour la convergence en probabilité, presque sûre ou en moyenne quadratique. Étant donné qu'il y a convergence vers une constante, rappelons (voir section 5.8) que la convergence en loi est équivalente à la convergence en probabilité. Pour $\Theta\subseteq\mathbb{R}^k$ la convergence en probabilité, donc la convergence en loi, et la convergence presque sûre s'entendent composante par composante. La convergence en moyenne quadratique se généralise avec la norme $\|.\|$ euclidienne usuelle dans $\mathbb{R}^k$.

Nous énonçons tout d'abord une propriété de convergence des moments empiriques de portée générale, dépassant le cadre paramétrique et que nous reprendrons donc dans le cadre non paramétrique du chapitre 8.

Proposition 6.1 *Si, pour la loi mère, $E(|X^r|)$ existe, alors tous les moments empiriques jusqu'à l'ordre r, simples ou centrés, sont des estimateurs presque sûrement convergents des moments correspondants de la loi.*

Il est clair que si les conditions d'application de la loi forte des grands nombres selon le théorème 5.3 sont réunies pour la v.a. X^r (r entier), alors le moment empirique M_r, comme moyenne des $X_1^r, X_2^r, ..., X_n^r$, converge presque sûrement vers μ_r, moyenne de la loi de X^r. Si nous nous en tenons à l'énoncé de ce théorème, la condition est que la variance de la loi considérée existe, donc, pour la loi de X^r, que $E(X^{2r})$ existe. Dans la proposition ci-dessus nous avons indiqué une condition plus faible qui résulte d'une version de la loi forte des grands nombres due à Kolmogorov.

Les moments d'ordres inférieurs existant a fortiori, ils convergent également. Quant au moment centré M'_k ($k \leq r$), il converge en tant que fonction continue de $M_1, M_2, ..., M_k$ et nécessairement vers μ'_k qui s'exprime par la même fonction vis-à-vis de $\mu_1, \mu_2, ..., \mu_k$ (voir la proposition 5.15 sur la convergence d'une fonction de v.a.).

En particulier si $E(X^2)$ existe ou, de façon équivalente, si la variance de la loi mère existe, la variance empirique $\widetilde{S}_n^2$ converge presque sûrement vers la variance de cette loi (et a fortiori $\bar{X}_n$ converge vers sa moyenne). Au passage notons que ceci vaut aussi pour la variance d'échantillon S_n^2 qui ne diffère de $\widetilde{S}_n^2$ que par le facteur $\frac{n}{n-1}$.

Proposition 6.2 *Soit une famille paramétrique de paramètre θ de dimension k telle que $E_\theta(|X^k|)$ existe pour tout θ et qu'il existe un estimateur des moments pour θ. Si les k premiers moments $\mu_1(\theta), ..., \mu_k(\theta)$ sont des fonctions continues de θ, alors cet estimateur est convergent presque sûrement.*

En effet en raison de l'hypothèse de continuité, la résolution du système d'équations de la définition 6.2 conduit à un estimateur des moments qui s'exprime comme une fonction continue, de $\mathbb{R}^k$ dans $\mathbb{R}^k$, des moments empiriques $\mu_1, \mu_2, ..., \mu_k$. En vertu de la proposition 5.15 il converge donc vers la solution du système $\mu_1(\theta) = \mu_1(\theta_0), \mu_2(\theta) = \mu_2(\theta_0), ..., \mu_k(\theta) = \mu_k(\theta_0)$ où nous distinguons ici θ_0 comme étant la vraie valeur de θ pour la loi mère (ainsi $M_r \xrightarrow{p.s.} \mu_r(\theta_0)$ pour $r = 1, ..., k$). Du fait de l'unicité de solution en θ pour ce système, propre à l'existence de l'estimateur des moments, cette solution ne peut être que θ_0.

La convergence est une condition sine qua non pour qualifier une statistique d'estimateur et elle sera normalement vérifiée pour les estimateurs naturels.

Pour la loi de Cauchy généralisée de paramètre θ définie par la densité :

$$f(x;\theta) = \frac{1}{\pi[1 + (x-\theta)^2]}, \quad x \in \mathbb{R},$$

(pour $\theta = 0$, c'est la loi de Student à 1 degré de liberté) on a vu dans l'exemple 2.1 que la moyenne n'existe pas. On peut se poser la question de savoir comment se comporte alors la moyenne empirique. On montre (via la fonction caractéristique des moments, comme proposé dans un exercice du chapitre 5)

que la moyenne $\bar{X}$ suit en fait la même loi ! Elle ne converge donc pas vers θ. En fait pour estimer θ il faut prendre la médiane de l'échantillon, laquelle est convergente.

6.5.4 Exhaustivité d'un estimateur

S'agissant d'estimer θ, certaines statistiques peuvent être exclues du fait qu'elles n'utilisent pas de façon exhaustive toute l'information contenue dans l'échantillon $X_1, X_2, ..., X_n$. A l'inverse on peut s'attendre à ce qu'un «bon» estimateur soit une statistique qui ne retienne que ce qui est utile de l'échantillon. Les notions d'exhaustivité et d'exhaustivité minimale viennent préciser cela. Dans cette section Θ pourra être de dimension quelconque tout comme les statistiques considérées.

Définition 6.8 *On dit que T_n est une* ***statistique exhaustive*** *pour $\theta \in \Theta \subseteq \mathbb{R}^k$ si la loi conditionnelle de $(X_1, X_2, ..., X_n)$ sachant T_n ne dépend pas de θ.*

Exemple 6.8 Soit la v.a. X de loi continue uniforme sur $[0, \theta]$ où θ est inconnu. Ecrivons sa fonction de densité sous la forme :

$$f(x;\theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$$

afin d'y intégrer le fait que le support de cette densité est $[0, \theta]$, lequel dépend donc de θ.

Pour un échantillon de taille n la densité conjointe est (en rappelant que nous la notons également par f pour simplifier, voir section 5.2) :

$$\begin{aligned} f(x_1, x_2, ..., x_n; \theta) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i) \\ &= \frac{1}{\theta^n} I_{[0,\infty[}(x_{(1)}) I_{]-\infty,\theta]}(x_{(n)}) \end{aligned}$$

où $x_{(1)} = \min\{x_1, x_2, ..., x_n\}$ et $x_{(n)} = \max\{x_1, x_2, ..., x_n\}$. Déterminons la densité conditionnelle de $X_1, X_2, ..., X_n$ sachant $X_{(n)} = t$ qui n'est définie que si $t \in [0\ , \theta]$. Par extension à plusieurs variables de l'expression vue en section 3.2, elle est égale au rapport de la densité conjointe de $(X_1, X_2, ..., X_n, X_{(n)})$ à la densité marginale de $X_{(n)}$. Or la densité conjointe de $(X_1, X_2, ..., X_n, X_{(n)})$ en un point quelconque $(x_1, x_2, ..., x_n, t)$ est égale à la densité conjointe de $(X_1, X_2, ..., X_n)$ en $(x_1, x_2, ..., x_n)$ si $t = \max\{x_1, x_2, ..., x_n\}$ et 0 sinon, ce qui peut s'exprimer par la densité conjointe de l'échantillon multipliée par un facteur $s(x_1, x_2, ..., x_n, t)$ valant 1 ou 0 indépendamment de θ. Par ailleurs nous

avons vu dans l'exemple 6.4 que la densité de $X_{(n)}$ au point $t \in [0, \theta]$ vaut nt^{n-1}/θ^n. D'où la densité conditionnelle de $X_1, X_2, ..., X_n$ sachant $X_{(n)} = t$:

$$\frac{(1/\theta^n) \,.\, I_{[0,+\infty[}(x_{(1)}) \,.\, I_{]-\infty,\theta]}(t) \,.\, s(x_1, x_2, ..., x_n, t)}{nt^{n-1}/\theta^n}$$

dans laquelle θ disparaît puisque, nécessairement, $t \in [0, \theta]$ ce qui entraîne $I_{]-\infty,\theta]}(t) = 1$.

La valeur maximale de l'échantillon est donc une statistique exhaustive pour l'estimation de θ. Intuitivement on sent bien que, θ devant être supérieur à toute observation, la valeur maximale observée dans l'échantillon livre toute l'information utile quant à la valeur possible de θ. Le même résultat peut être établi, et attendu intuitivement, pour la loi discrète uniforme sur $\{0, 1, 2, ..., r\}$. Supposons qu'on ait dans une urne r jetons numérotés de 1 à r où r est inconnu. On effectue n tirages (en principe avec remise) et l'on note les numéros $x_1, x_2, ..., x_n$ tirés. Il est clair que seul le numéro maximal observé est pertinent pour estimer r. ■

Dans cet exemple on a vu que le calcul de la densité conditionnelle est loin d'être immédiat. Le théorème suivant va nous simplifier la tâche.

Théorème 6.1 ***Théorème de factorisation.***
La statistique $T_n = t(X_1, X_2, \ldots, X_n)$ est exhaustive pour θ si et seulement si la densité de probabilité (ou fonction de probabilité) conjointe s'écrit, pour tout $(x_1, x_2, ..., x_n) \in \mathbb{R}^n$, sous la forme :

$$f(x_1, x_2, ..., x_n; \theta) = g(t(x_1, x_2, ..., x_n); \theta)\, h(x_1, x_2, ..., x_n).$$

Nous omettrons la démonstration de ce théorème que l'on trouvera dans des ouvrages plus avancés (avec d'ailleurs des conditions mineures de validité). Ce théorème indique que si, dans l'expression de la densité conjointe, θ entre uniquement dans un facteur contenant une certaine fonction de $x_1, x_2, ..., x_n$ alors cette fonction définit une statistique exhaustive. Notons, pour mémoire, que la notion d'exhaustivité et le théorème de factorisation reposent sur la densité conjointe seulement et, de ce fait, s'appliquent dans un cadre plus vaste que celui d'un échantillon aléatoire.

Reprenons l'exemple précédent où :

$$f(x_1, x_2, ..., x_n; \theta) = \frac{1}{\theta^n} I_{]-\infty,\theta]}(x_{(n)}).I_{[0,+\infty[}(x_{(1)}).$$

On voit immédiatement et sans calculs que $x_{(n)}$ forme avec θ un facteur isolé et, donc, que $X_{(n)}$ est exhaustive.

Exemple 6.9 Prenons le cas de la loi de Gauss $\mathcal{N}(\mu, \sigma^2)$ où le paramètre de dimension 2, (μ, σ^2), est inconnu. On a :

$$f(x_1, ..., x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}\}.$$

En développant $(x_i - \mu)^2$ et en regroupant les termes du produit on obtient :

$$f(x_1, ..., x_n; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^n x_i^2}{\sigma^2} - \frac{2\mu\sum_{i=1}^n x_i}{\sigma^2} + \frac{n\mu^2}{\sigma^2}\right)\right\}.$$

Comme n'apparaissent que $\sum_{i=1}^n x_i^2$ et $\sum_{i=1}^n x_i$, le couple $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ est une statistique exhaustive. Notons qu'ici, et c'est souvent le cas, dans la factorisation la fonction h est réduite à la constante 1. Les propositions ci-après montrent que $(\bar{X}_n, S_n^2)$ est également exhaustive. Ceci signifie que dans le cas où le phénomène étudié peut être considéré comme gaussien, on peut ne retenir de l'échantillon observé que sa moyenne et sa variance. Cette pratique est très répandue y compris en dehors du cadre gaussien ce qui peut signifier une perte d'information pour ce qui concerne l'estimation d'un paramètre inconnu. ■

Proposition 6.3 *Soit T_n une statistique exhaustive et T'_n une statistique telle que T_n soit une fonction de T'_n. Alors T'_n est également exhaustive.*

Pour montrer cela, explicitons les fonctions en jeu avec $T_n = u(T'_n)$ et $T'_n = t'(X_1, X_2, ..., X_n)$, d'où $T_n = u(t'(x_1, x_2, ..., x_n))$. Comme T_n est exhaustive la densité conjointe peut s'écrire :

$$f(x_1, x_2, ..., x_n; \theta) = g(u(t'(x_1, x_2, ..., x_n)); \theta)\, h(x_1, x_2, ..., x_n).$$

Le premier facteur contenant θ dépend des observations à travers la fonction $t'(x_1, x_2, ..., x_n)$ qui définit T'_n laquelle est donc exhaustive.

Proposition 6.4 *Soit T_n une statistique exhaustive et T'_n une statistique telle que $T'_n = r(T_n)$ où r est une fonction bijective. Alors T'_n est aussi exhaustive.*

Ceci résulte immédiatement du fait que l'on ait $T_n = r^{-1}(T'_n)$ et que l'on puisse appliquer la proposition 6.3.

La proposition 6.3 montre que la notion d'exhaustivité telle que nous l'avons définie n'implique pas une réduction au minimum de l'information utile dans l'échantillon pour estimer θ, mais une réduction suffisante (en anglais une statistique exhaustive est appelée *sufficient statistic*). Ainsi l'échantillon dans son ensemble : $(X_1, X_2, ..., X_n)$, est une statistique évidemment exhaustive. Or, s'il s'agit d'estimer un paramètre de dimension k, on peut s'attendre (à condition que le nombre d'observations n soit supérieur à k, ce que nous supposerons implicitement) à ce qu'une statistique exhaustive de dimension k procure un résumé minimal de l'information. Toutefois la proposition 6.4 nous dit que

celle-ci sera définie à une bijection près. Ainsi pour la famille $\mathcal{N}(\mu, \sigma^2)$, la statistique exhaustive $(\bar{X}_n, S_n^2)$ est sans doute minimale pour estimer (μ, σ^2). Nous pouvons définir formellement la notion d'exhaustivité minimale de la façon suivante.

Définition 6.9 *On dit que la statistique T_n^* est* ***exhaustive minimale*** *si elle est exhaustive et si, pour toute statistique exhaustive T_n, on peut trouver une fonction u telle que $T_n^* = u(T_n)$.*

La recherche d'une statistique exhaustive minimale ne sera pas abordée ici. Toutefois nous pourrons admettre intuitivement que, **si $\Theta \subseteq \mathbb{R}^k$, une statistique exhaustive à valeur dans $\mathbb{R}^k$ est en règle générale minimale.** La mise en évidence d'une statistique exhaustive minimale est particulièrement importante pour l'estimation. En effet une statistique qui contiendrait soit une partie seulement de l'information relative à θ, soit une part superflue, ne saurait être considérée comme un estimateur adéquat de θ. Nous énonçons donc le principe suivant : **tout estimateur pertinent est fonction d'une statistique exhaustive minimale.**

Pour ce qui concerne la classe exponentielle (voir section 6.3) montrons qu'une telle statistique existe et est aisément identifiable.

Proposition 6.5 *Soit une loi mère appartenant à une famille paramétrique de la classe exponentielle, avec un paramètre de dimension k. Alors, dans les notations de la définition 6.1, la statistique de dimension k :*

$$\left(\sum_{i=1}^{n} d_1(X_i), \sum_{i=1}^{n} d_2(X_i), ..., \sum_{i=1}^{n} d_k(X_i) \right)$$

est exhaustive minimale pour le paramètre inconnu.

Ceci résulte immédiatement du théorème de factorisation. En effet la densité (ou fonction de probabilité) conjointe $f(x_1, x_2, ..., x_n; \theta)$ peut s'écrire :

$$\prod_{i=1}^{n} f(x_i; \theta) = [a(\theta)]^n \prod_{i=1}^{n} b(x_i) \exp\left\{ c_1(\theta) \sum_{i=1}^{n} d_1(x_i) + \cdots + c_k(\theta) \sum_{i=1}^{n} d_k(x_i) \right\}.$$

Ainsi le tableau 6.1 nous livre directement les statistiques exhaustives minimales pour la plupart des lois usuelles.

Exemple 6.10 Soit la loi de Bernoulli $\mathcal{B}(p)$. On peut écrire sa fonction de probabilité :

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

$$f(x; \theta) = (1-p) \exp\{x . \ln \frac{p}{1-p}\},$$

qui répond à la forme de la classe exponentielle avec $d(x) = x$. On peut vérifier directement que $\sum_{i=1}^{n} X_i$ est exhaustive, car la densité conjointe :

$$\prod_{i=1}^{n} f(x_i;\theta) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

ne dépend que de $\sum_{i=1}^{n} x_i$. ■

Pour les lois binomiale, binomiale négative, de Poisson, exponentielle et gamma avec un seul paramètre inconnu on a également $d_1(x) = x$, c'est-à-dire que $\sum_{i=1}^{n} X_i$ ou, par bijection, $\bar{X}$ est une statistique exhaustive minimale. L'implication pratique de ce résultat est que les estimateurs pertinents du paramètre concerné sont à rechercher parmi les fonctions de $\bar{X}$ uniquement.

Pour la loi de Gauss le tableau 6.1 indique que le couple $(\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i)$ est exhaustif minimal ce qui corrobore le résultat trouvé de manière directe dans l'exemple 6.9.

Pour la loi de Pareto $d_1(x) = \ln x$, donc $\sum_{i=1}^{n} \ln X_i$ est exhaustive minimale et, celle-ci s'écrivant $\ln(\prod_{i=1}^{n} X_i)$, $\prod_{i=1}^{n} X_i$ l'est aussi.

Enfin pour la loi bêta le couple $(\sum_{i=1}^{n} \ln X_i\ ,\ \sum_{i=1}^{n} \ln(1-X_i))$ est exhaustif minimal pour (α, β), tout comme le couple $(\prod_{i=1}^{n} X_i\ , \prod_{i=1}^{n}(1-X_i))$.

On peut montrer que pour la classe des densités (ou fonctions de probabilité) répondant à certaines conditions dites **conditions de régularité** (précisées dans la note 6.1 ci-après) une famille de loi dont le paramètre inconnu est de dimension k ne peut admettre une statistique exhaustive dans $\mathbb{R}^k$ que si elle appartient à la classe exponentielle. **Il y a donc équivalence entre appartenance à la classe exponentielle et existence d'une statistique exhaustive de même dimension que le paramètre inconnu**.

Ainsi la famille des lois de Weibull, par exemple, qui répond aux conditions énoncées ci-après, mais qui n'est pas dans la classe exponentielle, n'admettra pas de statistique exhaustive de dimension 2. De fait, de par la forme de la densité :

$$f(x;\alpha,\lambda) = \alpha\lambda x^{\alpha-1}e^{-\lambda x^{\alpha}} \quad (x > 0),$$

aucune factorisation (au sens du théorème de factorisation) faisant apparaître une fonction de $x_1, x_2, ..., x_n$ à valeurs dans $\mathbb{R}^p$ avec $p < n$ n'est possible. Une statistique exhaustive minimale est donc de dimension n. Pour toute situation de ce type la statistique exhaustive minimale est, en fait, le vecteur des statistiques d'ordre $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ et non pas $(X_1, X_2, ..., X_n)$ lui-même. En tout état de cause aucun véritable résumé n'est possible si l'on veut conserver toute l'information utile pour estimer (λ, α). Ceci est également vrai pour le paramètre (α, β) de la loi de Gumbel et, dans le cas discret, pour le paramètre M de la loi hypergéométrique, ces lois n'appartenant pas à la classe exponentielle.

Note 6.1 *Conditions de régularité.*
Dans le cas d'une densité, des conditions suffisantes sont que $f(x;\theta)$ soit dérivable deux fois par rapport à x à l'intérieur du support de f et dérivable par rapport à θ dans Θ (ou par rapport à chacune de ses composantes si $k > 1$). Ceci est vérifié pour les familles classiques dont les densités reposent toutes sur des fonctions mathématiques dérivables. Mais **ceci exclut les familles dont le support dépend du paramètre inconnu**. Prenons de nouveau la loi $\mathcal{U}[0\ ,\theta]$ dont la densité est $f(x;\theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$. La présence de la fonction indicatrice empêche la dérivabilité par rapport à θ en tout $\theta \in \mathbb{R}^+$. En effet, x étant fixé, pour $\theta < x$ la densité vaut 0 et pour $\theta > x$ elle passe à $1/\theta$. Elle est donc discontinue et a fortiori non dérivable en $\theta = x$. C'est pourquoi cette famille, quoique n'étant pas dans la classe exponentielle, peut admettre toutefois une statistique exhaustive de dimension 1, à savoir le maximum des X_i comme on l'a montré dans l'exemple 6.8.

Pour conclure, la notion d'exhaustivité nous a permis de mettre en évidence, pour la plupart des lois usuelles, quelles sont les statistiques à retenir qui, à une fonction bijective près, devraient déboucher sur des estimateurs pertinents pour θ. Nous nous tournons maintenant vers la classe des estimateurs sans biais où nous pourrons encore préciser les choses et obtenir des résultats tangibles dans la recherche des meilleurs estimateurs.

6.6 Recherche des meilleurs estimateurs sans biais

6.6.1 Estimateurs UMVUE

Si l'on privilégie maintenant un estimateur sans biais l'objectif se ramène, pour le critère de l'erreur quadratique moyenne, à rechercher l'estimateur dont la variance, en l'occurrence la dispersion autour de θ lui-même, est minimale. Toutefois, comme θ est inconnu, cela n'a d'intérêt que si un tel estimateur domine tous les autres quel que soit $\theta \in \Theta$, c'est-à-dire *uniformément* en θ. Il est possible qu'un tel estimateur n'existe pas, mais nous allons voir dans le cas de la dimension 1 qu'il existe effectivement et peut être mis en évidence pour la classe des familles exponentielles qui recouvre la plupart des lois usuelles.

Définition 6.10 *On dit que l'estimateur T_n^* est **UMVUE** pour θ (uniformly minimum variance unbiased estimator) s'il est sans biais pour θ et si pour tout autre estimateur T_n sans biais on a :*

$$V_\theta(T_n^*) \leq V_\theta(T_n)\,, \quad \textit{pour tout } \theta \in \Theta\,.$$

Nous adoptons ici le sigle anglais UMVUE utilisé internationalement.

Proposition 6.6 *Si la famille de la loi mère appartient à la classe exponentielle avec paramètre de dimension 1 ($\Theta \subseteq \mathbb{R}$) et s'il existe une statistique*

fonction de la statistique exhaustive minimale $\sum_{i=1}^n d(X_i)$ *qui soit sans biais pour* θ*, alors elle est unique et elle est UMVUE pour* θ*.*

La démonstration de ce théorème dépasse le cadre de cet ouvrage. Selon cette proposition, pour trouver le meilleur estimateur, s'il existe, il suffit de **rechercher la fonction de** $\sum_{i=1}^n d(X_i)$ **qui soit sans biais pour** θ, quel que soit $\theta \in \Theta$. L'existence d'un estimateur UMVUE est donc subordonnée à celle d'une fonction de la statistique exhaustive minimale qui soit sans biais pour θ. Ce théorème vaut aussi pour estimer une fonction $h(\theta)$ de θ.

Exemple 6.11 Soit à estimer le paramètre $\lambda > 0$ de la loi exponentielle dont la densité est :

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad \text{si} \quad x \geq 0 \quad (0 \text{ sinon}).$$

Pour cette famille (voir tableau 6.1) on a $d(x) = x$ et la statistique exhaustive minimale est, à une bijection près, $\sum_{i=1}^n X_i$. Or $E(\sum_{i=1}^n X_i) = nE(X) = n/\lambda$ et cette statistique n'est évidemment pas sans biais pour λ. Examinons plutôt $n/\sum_{i=1}^n X_i = 1/\bar{X}$ dont on peut penser qu'elle fasse l'affaire puisque la moyenne de la loi est $1/\lambda$ (c'est l'estimateur par la méthode des moments). Calculons son espérance mathématique en posant $T_n = \sum_{i=1}^n X_i$. On sait (voir section 4.2.3) que T_n suit une loi $\Gamma(n\ ,\lambda)$ d'où, supposant $n > 1$:

$$\begin{aligned} E\left(\frac{1}{T_n}\right) &= \int_0^{+\infty} \frac{1}{t} \frac{\lambda}{(n-1)!} (\lambda t)^{n-1} e^{-\lambda t} dt \\ &= \frac{\lambda}{n-1} \int_0^{+\infty} \frac{\lambda}{(n-2)!} (\lambda t)^{n-2} e^{-\lambda t} dt \\ &= \frac{\lambda}{n-1}\ , \end{aligned}$$

car on reconnaît que l'expression intégrée est la densité de la loi $\Gamma(n-1\ ,\lambda)$. Donc, en fait, il faut choisir $(n-1)/T_n$ pour estimer λ sans biais, un résultat qui n'est absolument pas intuitif. Par conséquent $(n-1)/\sum_{i=1}^n X_i$ est l'estimateur UMVUE de λ. On ne peut en conclure directement qu'il domine en e.q.m. l'estimateur des moments, car celui-ci est biaisé. Toutefois, comme ce dernier vaut $n/(n-1)$ fois le premier sa variance est supérieure et il est effectivement dominé. ■

Pour la loi de Bernoulli la statistique exhaustive minimale est également $\sum_{i=1}^n X_i$ (voir tableau 6.1) et comme le paramètre p est la moyenne de la loi, $E(\bar{X}) = p$ et $\bar{X}$ est l'estimateur UMVUE. Dans une situation de sondage une proportion observée dans l'échantillon est UMVUE pour la proportion correspondante dans la population.

Pour la loi de Poisson $\bar{X}$, le nombre moyen d'occurrences observées par unité de temps (ou de surface dans un problème spatial), est également l'estimateur UMVUE pour λ.

Pour la loi de Gauss, à supposer que la variance σ^2 soit connue (une situation assez hypothétique mais souvent envisagée comme cas d'école) on a encore $d(x) = x$ pour ce qui concerne l'estimation de μ et $\bar{X}$ est donc UMVUE.

Le cas de la loi $\Gamma(r, \lambda)$, où seul λ est inconnu, est de même nature que celui de la loi $\mathcal{E}(\lambda)$.

Pour la loi de Pareto avec seuil a connu on a $d(x) = \ln x$ et l'on montre que $(n-1)/\sum_{i=1}^{n} \ln(X_i/a)$ est UMVUE pour θ (voir exercices).

Prenons encore un exemple utile en pratique, celui de la loi binomiale négative $\mathcal{BN}(r, p)$ où r est connu. En effet dans certaines situations on voudra estimer la probabilité de succès p dans un processus de Bernoulli en observant le nombre d'essais qu'il aura fallu effectuer jusqu'à voir le r-ième succès (nous prendrons ci-après la version de la loi binomiale négative où la v.a. est le nombre **total** d'essais, voir section 4.1.4).

Exemple 6.12 Soit X qui suit la loi $\mathcal{BN}(r, p)$ avec r connu et p inconnu, de fonction de probabilité :

$$f(x; p) = \binom{x-1}{x-r} p^r (1-p)^{x-r}, \quad x = r, r+1, ...$$

Elle appartient à la classe exponentielle avec $d(x) = x$ et $\sum_{i=1}^{n} X_i$ est donc statistique exhaustive minimale. Toutefois comme $E(X) = r/p$, elle n'est évidemment pas sans biais pour p. L'intuition nous oriente vers $r/\bar{X}$ (estimateur de la méthode des moments) mais le calcul montrerait que cette statistique reste biaisée. Nous allons voir qu'il faut prendre $(nr-1)/(\sum_{i=1}^{n} X_i - 1)$ pour obtenir la statistique sans biais et donc UMVUE.

Remarquons tout d'abord que $\sum_{i=1}^{n} X_i$ correspond au nombre d'essais jusqu'à atteindre le nr-ième succès, car il est licite de mettre les séquences d'essais bout à bout étant donné la nature du processus de Bernoulli. En conséquence cette statistique suit une loi $\mathcal{BN}(nr, p)$. Le problème à plusieurs observations est donc identique au problème à une seule observation : il suffit de remplacer r par nr. D'ailleurs, en pratique, on n'effectuerait qu'une série d'essais après avoir choisi une valeur de r. Restons-en donc à $n = 1$ et calculons

$$\begin{aligned} E\left(\frac{r-1}{X-1}\right) &= \sum_{x=r}^{\infty} \frac{r-1}{x-1} \frac{(x-1)!}{(x-r)!(r-1)!} p^r (1-p)^{x-r} \\ &= p \sum_{x=r}^{\infty} \frac{(x-2)!}{(x-r)!(r-2)!} p^{r-1} (1-p)^{x-r} \\ &= p \sum_{t=r-1}^{\infty} \frac{(t-1)!}{(t-(r-1))!(r-2)!} p^{r-1} (1-p)^{t-(r-1)} \end{aligned}$$

en posant $t = x - 1$. On reconnaît dans la dernière sommation le terme général de la loi $\mathcal{BN}(r-1\ , p)$, d'où $E\left((r-1)/(X-1)\right) = p$, ce qui prouve que

$(r-1)/(X-1)$ est l'estimateur recherché, un résultat que l'intuition ne pouvait laisser prévoir. ■

Si, sur le plan théorique, le cas des familles de la classe exponentielle est résolu, la recherche est plus délicate pour d'autres familles comme les lois uniforme $\mathcal{U}[0,\theta]$, Weibull, Gumbel et hypergéométrique. On montre cependant que pour les statistiques exhaustives minimales dites *complètes* - une propriété, hélas, généralement difficile à vérifier - la proposition 6.6 énoncée pour la classe exponentielle se généralise : une statistique exhaustive complète qui est sans biais est UMVUE pour θ. Par ailleurs un théorème (dit de Rao-Blackwell) établit, dans les conditions les plus générales, qu'à partir d'une statistique sans biais quelconque on peut déduire une statistique sans biais qui domine la première en la conditionnant sur une statistique exhaustive.

Ainsi, globalement, peut-on conclure qu'on aura toujours avantage à rechercher une fonction d'une statistique exhaustive, minimale si possible, qui soit sans biais pour θ.

Note 6.2 : Une statistique T est dite complète pour θ s'il n'existe pas de fonction de T, mis à part la fonction constante, dont l'espérance mathématique soit indépendante de θ, donc :

$$E_\theta[g(T)] = c \quad \text{pour tout} \quad \theta \in \Theta \;\Rightarrow\; g(t) = c \text{ pour toute valeur possible } t.$$

Bien que la définition ne concerne que l'espérance mathématique il s'ensuit, en vérité, qu'aucune fonction de T non constante ne peut avoir une loi indépendante de θ. Ceci n'est pas nécessairement vrai pour une statistique exhaustive minimale et le fait pour une statistique d'être complète signifie une réduction encore plus forte de l'information utile. Aussi une statistique complète est-elle a fortiori toujours exhaustive minimale.

A titre d'illustration montrons que $X_{(n)}$ est complète dans la famille $U[0\ ,\theta]$ et, comme $\frac{n+1}{n}X_{(n)}$ est sans biais pour θ (voir exemple 6.4), cette fonction de $X_{(n)}$ est donc nécessairement UMVUE. Nous avons vu dans l'exemple 6.4 que la densité de $X_{(n)}$ est $f(t;\theta) = nt^{n-1}\theta^{-n}$ pour $\theta \in \Theta$. Supposons qu'il existe $g(X_{(n)})$ telle que :

$$E[g(X_{(n)})] = \int_0^\theta g(t)nt^{n-1}\theta^{-n}dt = c \quad \text{pour tout} \quad \theta > 0$$

ou : $$\int_0^\theta [g(t)-c]nt^{n-1}\theta^{-n}dt = 0 \qquad \Longleftrightarrow \quad \int_0^\theta [g(t)-c]t^{n-1}dt = 0.$$

En dérivant par rapport à θ on obtient :

$$[g(\theta) - c]\theta^{n-1} = 0 \quad \text{pour tout} \quad \theta > 0\ ,$$

ce qui implique $g(\theta) = c$ et $X_{(n)}$ est complète. Ainsi, si l'on privilégie le choix d'un estimateur sans biais, $\frac{n+1}{n}X_{(n)}$ est celui qu'il faut retenir.

6.6.2 Estimation d'une fonction de θ et reparamétrisation

Comme nous en avons fait état en section 6.2, il se peut que l'on souhaite estimer une fonction $h(\theta)$ qui corresponde à une valeur caractéristique particulièrement intéressante de la loi mère. Si cette fonction est bijective et deux fois dérivable tous les résultats ci-dessus restent valables et le problème reste de rechercher une fonction d'une statistique exhaustive minimale qui soit sans biais pour $h(\theta)$. On peut aussi considérer $h(\theta)$ comme une reparamétrisation de la famille : posant $\rho = h(\theta)$ comme nouveau paramètre, il suffit de substituer $h^{-1}(\rho)$ à θ dans l'expression de $f(x;\theta)$.

Exemple 6.13 Soit à estimer $e^{-\lambda}$, la probabilité qu'il n'y ait aucune occurrence dans une unité de temps donnée, pour une loi de Poisson. Sachant que $\sum_{i=1}^{n} X_i = T$ est exhaustive minimale, montrons que $\left(\frac{n-1}{n}\right)^T$ est sans biais pour $e^{-\lambda}$, rappelant que T suit une loi $\mathcal{P}(n\lambda)$. D'où :

$$E\left(\left(\frac{n-1}{n}\right)^T\right) = \sum_{t=0}^{\infty}\left(\frac{n-1}{n}\right)^t \frac{e^{-n\lambda}(n\lambda)^t}{t!}$$
$$= e^{-n\lambda}\sum_{t=0}^{\infty}\frac{[(n-1)\lambda]^t}{t!}$$

qui est $e^{-\lambda}$ car la sommation est le développement en série entière de $e^{(n-1)\lambda}$. En conclusion $(\frac{n-1}{n})^{\sum_{i=1}^{n} X_i}$ est UMVUE pour $e^{-\lambda}$. ■

6.6.3 Borne de Cramer-Rao et estimateurs efficaces

Sous certaines conditions de régularité, à la fois pour la famille étudiée et pour l'estimateur sans biais considéré, on peut montrer que sa variance ne peut descendre au-dessous d'un certain seuil qui est fonction de θ. Ce seuil, appelé borne de Cramer-Rao, est intrinsèque à la forme de la densité (ou de la fonction de probabilité) $f(x;\theta)$. L'intérêt de ce résultat est que, si l'on trouve un estimateur sans biais dont la variance atteint ce seuil, alors il est le meilleur possible (UMVUE) parmi les estimateurs sans biais «réguliers».

Théorème 6.2 *(**Inégalité de Cramer-Rao** ou de Fréchet). Soit T un estimateur sans biais pour θ de dimension 1. Sous certaines conditions de régularité on a nécessairement, pour tout $\theta \in \Theta$:*

$$V_\theta(T) \geq \frac{1}{nI(\theta)},$$

*où $I(\theta)$, appelé **information de Fisher**, vaut :*

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

Nous omettrons la démonstration de ce théorème.

Si un estimateur sans biais pour θ atteint la borne de Cramer-Rao, on dit qu'il est *efficace*.

Note 6.3 Les conditions de régularité, dans le cas continu, sont les suivantes :

- $I(\theta)$ existe pour tout $\theta \in \Theta$
- la dérivée par rapport à θ d'une intégrale sur la densité conjointe

$$\int \cdots \int f(x_1, x_2, ..., x_n; \theta) dx_1 dx_2 \cdots dx_n$$

peut s'obtenir en dérivant à l'intérieur de l'intégrale

- la dérivée par rapport à θ de $E_\theta(T)$ peut s'obtenir en dérivant à l'intérieur de l'intégrale correspondante
- le support de $f(x;\theta)$ est indépendant de θ.

Dans le cas discret les conditions portent sur les sommations en lieu et place des intégrations.

Avant d'illustrer cette inégalité de Cramer-Rao montrons succinctement que $I(\theta)$ peut aussi se calculer selon :

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

ce qui facilitera généralement les calculs (toutefois cela suppose bien sûr que cette expression existe, mais aussi que l'on puisse dériver deux fois sous le signe somme comme dans la démonstration qui suit). Nous nous restreindrons au cas où f est une densité.

Démonstration. Posons :

$$U = \frac{\partial}{\partial \theta} \ln f(X;\theta) = \frac{\frac{\partial}{\partial \theta} f(X;\theta)}{f(X;\theta)}.$$

On a :

$$E_\theta(U) = \int_{\mathbb{R}} \frac{\frac{\partial}{\partial \theta} f(x;\theta)}{f(x;\theta)} f(x;\theta) dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x;\theta) dx = \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x;\theta) dx = 0$$

puisque cette intégrale est égale à la constante 1. De plus :

$$\begin{aligned} \frac{\partial^2}{\partial \theta^2} \ln f(x;\theta) &= \frac{\frac{\partial^2}{\partial \theta^2} f(x;\theta).f(x;\theta) - [\frac{\partial}{\partial \theta} f(x;\theta)]^2}{[f(x;\theta]^2} \\ &= \frac{\frac{\partial^2}{\partial \theta^2} f(x;\theta)}{f(x;\theta)} - \left[\frac{\frac{\partial}{\partial \theta} f(x;\theta}{f(x;\theta)} \right]^2, \end{aligned}$$

d'où :

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

Or :

$$E_\theta\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right] = \int_{\mathbb{R}}\frac{\partial^2}{\partial\theta^2}f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}}f(x;\theta)dx = 0,$$

ce qui démontre la relation. □

L'expression de l'information de Fisher pour certaines lois de la classe exponentielle est donnée dans le **tableau 6.1** de la section 6.3.

Exemple 6.14 Soit à estimer le paramètre λ dans la famille des lois $\mathcal{E}(\lambda)$ de densités $f(x;\lambda) = \lambda e^{-\lambda x}$ pour $x \geq 0$. Déterminons la borne de Cramer-Rao. On a :

$$\begin{aligned}\ln f(x;\lambda) &= \ln\lambda - \lambda x\\ \frac{\partial}{\partial\lambda}\ln f(x;\lambda) &= \frac{1}{\lambda} - x,\\ \frac{\partial^2}{\partial\lambda^2}\ln f(x;\lambda) &= -\frac{1}{\lambda^2}.\end{aligned}$$

D'où :

$$I(\lambda) = E\left(\frac{1}{\lambda^2}\right) = \frac{1}{\lambda^2}$$

et la borne de Cramer-Rao est donc égale à λ^2/n.

Dans l'exemple 6.11 on a vu que $(n-1)/\sum_{i=1}^n X_i$ est UMVUE pour λ. Par le même type d'argument que pour le calcul de l'espérance de cet estimateur effectué alors dans cet exemple, on montre (en supposant $n > 2$) que sa variance est :

$$V\left(\frac{n-1}{\sum_{i=1}^n X_i}\right) = \frac{\lambda^2}{n-2} > \frac{\lambda^2}{n}.$$

Comme il s'agit là du meilleur estimateur possible, la borne de Cramer-Rao n'est donc pas atteignable : il n'existe pas d'estimateur efficace pour λ. ■

Le problème qui se pose est de savoir si et quand il existe un estimateur sans biais pour θ, ou éventuellement pour une fonction $h(\theta)$ qui atteigne la borne de Cramer-Rao. La proposition 6.7 apportera la réponse, mais pour bien la comprendre il n'est pas inutile de voir au préalable les implications d'une

reparamétrisation de la famille étudiée. Considérons donc un changement de paramètre $\rho = h(\theta)$ qui ne remette pas en cause les conditions de régularité. Pour un estimateur sans biais de ρ la variance est bornée par $\frac{1}{nI(\rho)}$. Montrons que $I(\rho)$ peut se déduire aisément de $I(\theta)$.

En notant simplement par $f(x;\rho)$ la densité (ou fonction de probabilité) reparamétrisée avec ρ, on a :

$$\frac{\partial}{\partial\rho}\ln f(x;\rho) = \frac{\partial}{\partial\theta}\ln f(x;\theta)\,\frac{d\theta}{d\rho} = \frac{\partial}{\partial\theta}\ln f(x;\theta)\,\frac{1}{h'(\theta)}\,,$$

d'où :

$$I(\rho) = E_\rho\left[\left(\frac{\partial}{\partial\rho}\ln f(X;\rho)\right)^2\right] = \frac{1}{[h'(\theta)]^2}E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = \frac{I(\theta)}{[h'(\theta)]^2},$$

où θ doit être remplacé par $h^{-1}(\rho)$.

En d'autres termes, la borne de Cramer-Rao pour estimer $h(\theta)$ est :

$$\frac{[h'(\theta)]^2}{nI(\theta)}.$$

Proposition 6.7 *La borne de Cramer-Rao n'est atteinte que :*
- si la famille de lois est dans la classe exponentielle
- et pour l'estimation d'une fonction de reparamétrisation particulière de θ, *à savoir* $h(\theta) = E_\theta\left(\sum_{i=1}^n d(X_i)\right)$.

Nous admettrons cette proposition. Ainsi il n'existe qu'une fonction de θ qui puisse être estimée de façon «efficace». Pour déterminer cette fonction il suffit de calculer l'espérance mathématique de $\sum_{i=1}^n d(X_i)$ qui en est donc l'estimateur efficace. En réalité, pour être plus précis, cette fonction est définie à une transformation linéaire près, $ah(\theta)+b$ étant estimé sans biais par $a\sum_{i=1}^n d(X_i) + b$.

Cette proposition montre que, malgré tout, le résultat de Cramer-Rao est d'un intérêt limité.

Exemple 6.15 (suite de l'exemple 6.14) Pour la loi exponentielle la borne de Cramer-Rao ne peut être atteinte que pour estimer la fonction de λ pour laquelle $\sum_{i=1}^n d(X_i) = \sum_{i=1}^n X_i$ est sans biais, à savoir $E(\sum_{i=1}^n X_i) = nE(X) = \frac{n}{\lambda}$. En reparamétrant avec $\theta = h(\lambda) = \frac{1}{\lambda}$, θ est alors la moyenne de la loi et $\bar{X}$ est l'estimateur qui atteint la borne de Cramer-Rao pour θ. Celle-ci est :

$$\frac{[h'(\lambda)]^2}{nI(\lambda)} = \frac{(-1/\lambda^2)^2}{n(1/\lambda^2)} = \frac{1}{n\lambda^2} = \frac{\theta^2}{n}.$$

On vérifie directement que $V(\bar{X}) = \frac{1}{n}V(X) = \frac{\theta^2}{n}$ puisque $\theta^2 = \frac{1}{\lambda^2}$ est la variance de la loi. ■

Pour la loi de Bernoulli et la loi de Poisson où l'on a également $d(x) = x$, la moyenne $\bar{X}$ estime «efficacement» les paramètres respectifs p et λ qui sont les moyennes théoriques.

Pour la loi $\mathcal{BN}(r, p)$ avec r connu, on a également $d(x) = x$. La fonction de p qui est estimée efficacement est donc $E(\sum_{i=1}^{n} X_i) = \frac{nr(1-p)}{p}$ ou, plus simplement, $E(\bar{X}) = \frac{r(1-p)}{p}$.

Pour la loi de Pareto (a connu, θ inconnu), $d(x) = \ln x$ et $\sum_{i=1}^{n} \ln X_i$ estime sans biais $\frac{n}{\theta} + n \ln a$ (voir exercices) ou, de façon équivalente, $\frac{1}{n} \sum_{i=1}^{n} \ln(\frac{X_i}{a})$ estime sans biais $\frac{1}{\theta}$, de façon efficace.

Ces deux derniers cas illustrent l'importance très relative de la notion d'efficacité dans la mesure où elle est réalisée pour des fonctions du paramètre ne présentant pas nécessairement un intérêt central.

6.6.4 Extension à un paramètre de dimension $k > 1$

Evacuons tout d'abord le cas où l'on s'intéresse à une fonction $h(\theta)$ à valeurs dans $\mathbb{R}$. En effet les notions de biais, de convergence, d'erreur quadratique, d'estimateur UMVUE restent valables. Il faut cependant noter que les qualités de sans biais ou de variance minimale doivent être vérifiées pour tout $\theta \in \Theta \subseteq \mathbb{R}^k$, ce qui peut poser problème. On peut, par exemple, vouloir estimer le quantile d'ordre 0,95 de la loi $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu, soit la fonction $g(\mu, \sigma^2) = \mu + 1,645\sqrt{\sigma^2}$. Les qualités d'un estimateur doivent alors être examinées quel que soit le couple (μ, σ^2).

Nous considérons ici l'**estimation simultanée** de toutes les composantes $\theta_1, \theta_2, ..., \theta_k$ de θ. Le critère d'exhaustivité a déjà été traité avec $k > 1$. De même la généralisation de la notion de convergence a été évoquée en section 6.5.3. Un estimateur $T = (T_1, T_2, ..., T_k)$ étant un vecteur aléatoire dans $\mathbb{R}^k$ le biais est naturellement défini par le vecteur $\mathbb{E}_\theta(T) - \theta$ de composantes $E_\theta(T_1) - \theta_1, ..., E_\theta(T_k) - \theta_k$ (voir l'espérance mathématique d'un vecteur aléatoire en section 3.8). Pour ce qui concerne l'extension de la notion de variance nous pouvons prendre la somme des variances des composantes $\sum_{j=1}^{k} V_\theta(T_j)$, le critère d'écart quadratique correspondant étant $\|T - \theta\|^2$ où $\|.\|$ est la norme euclidienne usuelle dans $\mathbb{R}^k$. En effet l'e.q.m. devient :

$$E_\theta(\|T - \theta\|^2) = E_\theta\left(\sum_{j=1}^{k}(T_j - \theta_j)^2\right) = \sum_{j=1}^{k} E_\theta([T_j - \theta_j]^2)$$

$$= \sum_{j=1}^{k}(E_\theta(T_j) - \theta_j)^2 + \sum_{j=1}^{k} V_\theta(T_j)$$

où le premier terme est le carré de la norme du vecteur des biais et le deuxième la variance globale retenue.

Toutefois ce critère présente deux inconvénients majeurs. Le premier est qu'il est sensible aux différences d'échelle entre les composantes. Ceci peut être atténué en introduisant des pondérations de réduction de ces échelles. Le deuxième est qu'il ne tient pas compte des covariances existant généralement entre les composantes T_j du fait que ces T_j sont des statistiques fonctions des mêmes observations $X_1, X_2, ..., X_n$.

Il est donc préférable d'utiliser une notion de dispersion fondée sur la matrice des variances-covariances $\mathbb{V}_\theta(T)$ du vecteur aléatoire T (voir section 3.8). La plus répandue est celle du déterminant de cette matrice, lequel mesure à un facteur de proportionnalité près le volume d'ellipsoïdes de concentration autour du point moyen $\mathbb{E}_\theta(T)$ dans $\mathbb{R}^k$. Avec ce critère on a un résultat analogue à la proposition 6.6 : si la famille appartient à la classe exponentielle et s'il existe un estimateur fonction du vecteur des statistiques exhaustives minimales

$$\left(\sum_{i=1}^{n} d_1(X_i), \sum_{i=1}^{n} d_2(X_i), ..., \sum_{i=1}^{n} d_k(X_i) \right)$$

qui soit sans biais pour θ, alors il minimise ce critère uniformément en θ parmi les estimateurs sans biais.

En fait il minimise également le critère de la somme simple (et même pondérée) des variances $\sum_{j=1}^{k} V_\theta(T_j)$. De plus il fournit l'estimateur UMVUE pour chaque composante θ_j prise isolément. En règle générale on aura avantage, comme pour $k = 1$, à rechercher une statistique vectorielle qui soit fonction des statistiques exhaustives minimales et qui soit sans biais pour θ.

Exemple 6.16 Soit la famille $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu, appartenant à la classe exponentielle. La statistique $(\sum_{i=1}^{n} X_i\,, \sum_{i=1}^{n} X_i^2)$ est exhaustive minimale tout comme $(\bar{X}, S^2)$, laquelle est sans biais pour (μ, σ^2) (voir exemple 6.9). $(\bar{X}, S^2)$ est donc l'estimateur qui, parmi tous les estimateurs sans biais, a le déterminant de sa matrice des variances-covariances le plus faible, quel que soit (μ, σ^2). On a vu au chapitre 5 que $V(\bar{X}) = \sigma^2/n$, $V(S^2) = 2\sigma^4/(n-1)$ ainsi que l'indépendance de $\bar{X}$ et S^2. Ce déterminant est donc :

$$\det \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix} = \frac{2\sigma^6}{n(n-1)}.$$

De plus $\bar{X}$ est UMVUE pour μ et S^2 est UMVUE pour σ^2. ■

Exemple 6.17 Soit la famille $\mathcal{U}[a, b]$, où (a, b) est inconnu, qui n'appartient pas à la classe exponentielle. On démontre que $(X_{(1)}, X_{(n)})$ est exhaustive minimale (et complète). A partir des densités de $X_{(1)}$ et $X_{(n)}$ on établit aisément

les deux équations suivantes :

$$\begin{cases} E(X_{(1)}) &= a + \dfrac{b-a}{n+1} \\ \\ E(X_{(n)}) &= b - \dfrac{b-a}{n+1} \end{cases} .$$

Pour trouver la fonction dans $\mathbb{R}^2$ sans biais pour (a, b) il suffit de résoudre ce système d'équations en a et b. La solution étant

$$\begin{cases} a &= \dfrac{nE(X_{(1)}) - E(X_{(n)})}{n-1} \\ \\ b &= \dfrac{nE(X_{(n)}) - E(X_{(1)})}{n-1} \end{cases}$$

le couple $\left(\frac{nX_{(1)}-X_{(n)}}{n-1}, \frac{nX_{(n)}-X_{(1)}}{n-1}\right)$ est sans biais pour (a, b) et donc optimal au même sens que dans l'exemple précédent. ■

Pour finir décrivons la généralisation de l'inégalité de Cramer-Rao pour $k > 1$.

On introduit *la matrice d'information* $\mathbb{I}(\theta)$ symétrique d'ordre k dont l'élément en position (i, j) est :

$$E_\theta \left[\frac{\partial}{\partial \theta_i} \ln f(X;\theta) \; \frac{\partial}{\partial \theta_j} \ln f(X;\theta) \right] ,$$

moyennant les mêmes types de conditions de régularité que pour $k = 1$. On montre que cet élément peut aussi se calculer par $-E_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f(X;\theta) \right]$. Alors, pour tout estimateur sans biais T, la variance de toute combinaison linéaire $u^t T$ des composantes de T, où u est un vecteur quelconque de $\mathbb{R}^k$, reste supérieure ou égale à $u^t \frac{[\mathbb{I}(\theta)]^{-1}}{n} u$. Sachant que $V_\theta(u^t T) = u^t \mathbb{V}_\theta(T) u$ où $\mathbb{V}_\theta(T)$ est la matrice des variances-covariances de la statistique T (voir section 3.8), il est équivalent de dire que $\mathbb{V}_\theta(T) - \frac{1}{n}[\mathbb{I}(\theta)]^{-1}$ est une matrice semi-définie positive, ce que l'on note $\mathbb{V}_\theta(T) \geq \frac{1}{n}[\mathbb{I}(\theta)]^{-1}$, quel que soit θ.

Exemple 6.18 Prenons le cas de la loi $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu. On a, en posant $v = \sigma^2$:

$$f(x;\mu,v) = \frac{1}{\sqrt{2\pi v}} \exp\left\{ -\frac{1}{2v}(x-\mu)^2 \right\}$$

$$\ln f(x;\mu,v) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln v - \frac{1}{2v}(x-\mu)^2$$

$$\frac{\partial}{\partial \mu} \ln f(x;\mu,v) = \frac{1}{v}(x-\mu) \qquad \frac{\partial}{\partial v} \ln f(x;\mu,v) = -\frac{1}{2v} + \frac{(x-\mu)^2}{2v^2} .$$

En position (1,1) de la matrice $\mathbb{I}(\mu, \sigma^2)$ on trouve :

$$E\left[\frac{1}{v^2}(X-\mu)^2\right] = \frac{1}{v^2}v = \frac{1}{v} = \frac{1}{\sigma^2}\,,$$

et en position (2,2) :

$$\begin{aligned} E\left[\left(-\frac{1}{2v} + \frac{(X-\mu)^2}{2v^2}\right)^2\right] &= E\left[\frac{1}{4v^2} - \frac{(X-\mu)^2}{2v^3} + \frac{(X-\mu)^4}{4v^4}\right] \\ &= \frac{1}{4v^2} - \frac{v}{2v^3} + \frac{3v^2}{4v^4} = \frac{1}{2v^2} = \frac{1}{2\sigma^4} \end{aligned}$$

sachant que $E\left[(X-\mu)^4\right] = 3\sigma^4$. En position (1,2) ou (2,1) on a :

$$E\left[\frac{1}{v}(X-\mu)(-\frac{1}{2v} + \frac{(X-\mu)^2}{2v^2})\right] = 0$$

car cette expression ne contient que des moments centrés d'ordre impair. D'où :

$$\mathbb{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \quad \text{et} \quad \frac{\left[\mathbb{I}(\mu, \sigma^2)\right]^{-1}}{n} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

Cette matrice est précisément la matrice des variances-covariances du couple $(\bar{X}, S^2)$ qui est donc non seulement optimal au sens vu dans l'exemple 6.16 mais, de plus, «efficace» pour estimer (μ, σ^2). ■

6.7 L'estimation par la méthode du maximum de vraisemblance

Nous abordons maintenant deux méthodes générales qui, comme la méthode des moments vue en section 6.4, apportent des solutions dans des situations variées : l'approche par le maximum de vraisemblance et l'approche bayésienne. Nous commençons par celle du maximum de vraisemblance qui est la plus universelle (y compris pour des modèles complexes) pour deux raisons :

1. Elle est facile à mettre en oeuvre, se ramenant à un problème classique de résolution numérique.
2. Elle est optimale et même «efficace» *asymptotiquement*, i.e. quand la taille de l'échantillon tend vers l'infini. D'un point de vue pratique, pour un échantillon suffisamment grand (disons $n > 30$ pour fixer les idées), elle fournit des estimateurs de très bonne qualité.

6.7.1 Définitions

Définition 6.11 *Soit un échantillon aléatoire $(X_1, X_2, ..., X_n)$ dont la loi mère appartient à une famille paramétrique de densités (ou fonctions de probabilité) $\{f(x;\theta)\ , \theta \in \Theta\}$ où $\Theta \subseteq \mathbb{R}^k$. On appelle* ***fonction de vraisemblance*** *de θ pour une réalisation donnée $(x_1, x_2, ..., x_n)$ de l'échantillon, la fonction de θ :*

$$L(\theta; x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta).$$

L'expression de la fonction de vraisemblance est donc la même que celle de la densité (ou fonction de probabilité) conjointe mais le point de vue est différent. Ici les valeurs $x_1, x_2, ..., x_n$ sont fixées (ce seront les valeurs effectivement observées) et on s'intéresse à la façon dont varie la valeur de la densité (ou fonction de probabilité) associée à une série d'observations donnée suivant les différentes valeurs de θ. Dans le cas discret il s'agit directement de la probabilité $P_\theta(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$. S'il n'y a pas d'ambiguité possible, on notera la fonction de vraisemblance simplement $L(\theta)$. On dira que la valeur θ_1 de θ est *plus vraisemblable* que la valeur θ_2 si $L(\theta_1) > L(\theta_2)$. En ce sens il devient naturel de choisir pour θ la valeur la plus vraisemblable, disons $\widehat{\theta}^{MV}$, c'est-à-dire telle que la loi $f(x; \widehat{\theta}^{MV})$ correspondante confère la plus forte probabilité (ou densité de probabilité) aux observations relevées.

Définition 6.12 *On appelle* ***estimation du maximum de vraisemblance*** *une valeur $\hat{\theta}^{MV}$, s'il en existe une, telle que :*

$$L(\hat{\theta}^{MV}) = \sup_{\theta \in \Theta} L(\theta).$$

Une telle solution est fonction de $(x_1, x_2, ..., x_n)$, soit $\hat{\theta}^{MV} = h(x_1, x_2, ..., x_n)$. Cette fonction h induit la statistique (notée abusivement, mais commodément, avec le même symbole que l'estimation) $\hat{\theta}^{MV} = h(X_1, X_2, ..., X_n)$ appelée ***estimateur du maximum de vraisemblance (EMV)****.*

Cette définition appelle quelques remarques :

- $\hat{\theta}^{MV}$ est une fonction de $\mathbb{R}^n$ dans $\mathbb{R}^k$, associant à tout échantillon particulier une valeur particulière de θ ;
- généralement l'EMV existe et il est unique, i.e. quel que soit $(x_1, x_2, ..., x_n)$ il y a un et un seul maximum pour $L(\theta)$. On verra cependant dans l'exemple 6.22 un cas où il y a plusieurs solutions ;
- la définition de l'EMV s'étend à des variables aléatoires non i.i.d. car elle ne repose que sur la notion de densité (fonction de probabilité) conjointe. Elle s'étend même dans un cadre non paramétrique (voir section 8.5.3) ;

– une fois la réalisation $(x_1, x_2, ..., x_n)$ observée, l'estimation est facilement obtenue, y compris pour des situations complexes. Il suffit d'utiliser un algorithme de maximisation numérique comme on en trouve dans tous les logiciels mathématiques.

Quand les densités (fonctions de probabilité) conjointes sont des produits de fonctions puissances et exponentielles, ce qui est le cas la plupart du temps, on a plutôt intérêt à maximiser $\ln L(\theta)$, appelée *log-vraisemblance,* ce qui est équivalent puisque la fonction logarithmique est strictement croissante. Dans les cas «réguliers» où $L(\theta)$ est continûment dérivable et le support pour la famille de lois considérée est indépendant de θ, l'estimation par le maximum de vraisemblance (MV) vérifie (pour $\Theta \subseteq \mathbb{R}$) :

$$\frac{\partial}{\partial\theta} \ln L(\theta) = 0$$

$$\text{ou} \qquad \frac{\partial}{\partial\theta} \ln \left[\prod_{i=1}^{n} f(x_i, \theta) \right] = 0$$

$$\text{ou} \qquad \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(x_i, \theta) = 0.$$

Cette dernière égalité s'appelle *l'équation de vraisemblance*. Dans le cas où θ possède k dimensions $(\theta_1, \theta_2, ..., \theta_k)$, on résout un système de k équations obtenues en dérivant par rapport à chacune des composantes. Mathématiquement, le fait d'être solution de l'équation (ou du système d'équations) de vraisemblance n'est pas une condition suffisante pour être un maximum. Toutefois étant donné que $L(\theta)$ admet une borne supérieure en tant que probabilité (cas discret) mais aussi, généralement, en tant que densité de probabilité (cas continu), et qu'elle est le plus souvent concave, l'équation admettra une solution unique qui sera alors nécessairement un maximum. Dans les exemples, pour alléger l'exposé, nous n'examinerons pas dans le détail si la solution de l'équation (ou du système d'équations) correspond effectivement à un maximum.

6.7.2 Exemples et propriétés

Exemple 6.19 Soit la famille de Pareto où $a > 0$ est connu et θ est inconnu. On a :

$$f(x;\theta) = \theta a^{\theta} x^{-(\theta+1)} \ , \quad \text{si} \ \ x \geq a \ \ \text{et} \ \ \theta > 0$$

$$\ln f(x;\theta) = \ln\theta + \theta \ln a - (\theta+1) \ln x$$

$$\frac{\partial}{\partial\theta} \ln f(x;\theta) = \frac{1}{\theta} + \ln a - \ln x.$$

L'équation de vraisemblance s'écrit :

$$\frac{n}{\theta} + n \ln a - \sum_{i=1}^{n} \ln x_i = 0 \quad \text{ou} \quad \frac{n}{\theta} - \sum_{i=1}^{n} \ln(\frac{x_i}{a}) = 0$$

d'où l'estimation : $\hat{\theta}^{MV} = n \left[\sum_{i=1}^{n} \ln(\frac{x_i}{a})\right]^{-1}$

et l'EMV : $\hat{\theta}^{MV} = n \left[\sum_{i=1}^{n} \ln(\frac{X_i}{a})\right]^{-1}$. ■

Exemple 6.20 Soit la famille $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu. On a, en posant $v = \sigma^2$,

$$f(x, \mu, v) = \frac{1}{\sqrt{2\pi v}} \exp\{-\frac{1}{2v}(x-\mu)^2\}$$

$$\ln f(x; \mu, v) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln v - \frac{1}{2v}(x-\mu)^2 .$$

En dérivant par rapport à μ d'une part et par rapport à v d'autre part, puis en remplaçant x par x_i et en sommant sur $i = 1, ..., n$, on obtient le système d'équations de vraisemblance :

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{x_i - \mu}{v}\right) = 0 \\ \displaystyle -\frac{n}{v} + \frac{1}{v^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

d'où la solution $\hat{\mu} = \bar{x}$ et $\hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. L'EMV du paramètre (μ, σ^2) est donc $(\bar{X}, \widetilde{S}^2)$. ■

On constate sur ce dernier exemple que l'EMV peut avoir un biais. Dans les deux exemples il ne dépend que des statistiques exhaustives minimales, ce qui est une propriété générale.

Proposition 6.8 *Si T est une statistique exhaustive pour θ alors $\hat{\theta}^{MV}$, s'il existe, est fonction de T.*

Ceci résulte immédiatement du théorème de factorisation. Si la statistique $T = t(X_1, X_2, ..., X_n)$ est exhaustive, alors la densité (fonction de probabilité) conjointe est de la forme :

$$g(t(x_1, x_2, ..., x_n), \theta)\ h(x_1, x_2, ..., x_n) .$$

La maximisation vis-à-vis de θ ne concerne que la fonction g et la solution ne dépend donc des observations qu'à travers la fonction t. Cette proposition est vraie pour toute statistique exhaustive et en particulier pour une statistique exhaustive minimale. Remarquons que, bien que très intéressante, cette propriété n'entraîne pas que l'EMV soit UMVUE car, nous l'avons vu, il peut être biaisé.

La proposition suivante, que nous admettrons, montre l'intérêt de l'EMV dans le cas où il existe un estimateur efficace (pour $\Theta \subseteq \mathbb{R}$).

Proposition 6.9 *Si la famille de lois considérée répond à certaines conditions de régularité et si elle admet un estimateur sans biais efficace pour θ, alors l'EMV existe et est cet estimateur.*

Les conditions sont analogues à celles du théorème 6.2 auxquelles s'ajoute le fait que $L(\theta)$ admette une dérivée seconde continue. On pourra vérifier que l'EMV est $\bar{X}$ pour le paramètre λ de la loi de Poisson et pour le paramètre p de la loi de Bernoulli. Pour la loi exponentielle cela est vérifié pour la reparamétrisation $\theta = 1/\lambda$. Prenons maintenant deux exemples de cas non réguliers.

Exemple 6.21 Soit la famille de loi $\mathcal{U}[0, \theta]$. Reprenant l'exemple 6.8 on voit que la fonction de vraisemblance est :

$$L(\theta) = \frac{1}{\theta^n} I_{[0,+\infty[}(x_{(1)}).I_{]-\infty,\theta]}(x_{(n)})\,.$$

Elle contient $1/\theta^n$ qui est une fonction décroissante de θ, mais à partir du moment où $I_{]-\infty,\theta]}(x_{(n)}) = 1$, c'est-à-dire $\theta \geq x_{(n)}$. Par conséquent, le maximum est atteint pour $\theta = x_{(n)}$, puisque pour $\theta < x_{(n)}$ la fonction de vraisemblance est nulle. L'EMV est donc $X_{(n)}$. Nous avons vu (voir note 6.2) que l'estimateur UMVUE est $\frac{n+1}{n} X_{(n)}$. L'EMV en est proche, mais il est légèrement biaisé (si n n'est pas trop petit). Il est beaucoup plus pertinent que celui de la méthode des moments obtenu par $\bar{X} = \frac{\hat{\theta}^M}{2}$, soit $\hat{\theta}^M = 2\bar{X}$, qui ne repose pas sur une statistique exhaustive et est intuitivement peu convaincant (voir exercices). ■

Exemple 6.22 Considérons la loi de Laplace, ou loi exponentielle double (voir exercices du chapitre 2), de densité :

$$f(x; \mu) = \frac{1}{2} e^{-|x-\mu|} \quad \text{pour} \quad x \in \mathbb{R}\,.$$

Nous ne sommes pas ici dans un cas régulier car cette densité n'est pas dérivable quand $x = \mu$. La fonction de vraisemblance :

$$L(\mu) = \frac{1}{2^n} e^{-\sum_{i=1}^{n} |x_i - \mu|}$$

n'est donc pas dérivable par rapport à μ pour $\mu = x_1, \mu = x_2, ..., \mu = x_n$. Considérons la log-vraisemblance :

$$\ln L(\mu) = -n \ln 2 - \sum_{i=1}^{n} |x_i - \mu|.$$

Elle est maximale quand $\sum_{i=1}^{n} |x_i - \mu|$ est minimale. La dérivée de $|x_i - \mu|$, pour $\mu \neq x_i$, est égale au signe de $(x_i - \mu)$. On peut donc annuler la dérivée de $\ln L(\mu)$, si n est pair, en prenant pour valeur de μ une médiane de l'échantillon, soit tout point dans l'intervalle $(x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)})$. Admettant qu'on atteigne bien

ainsi un minimum (ce qui est intuitif, la médiane étant au «centre» des observations) on voit que l'estimateur du MV n'est pas unique. Dans le cas où n est impair la solution reste la médiane, qui est alors unique. La méthode des moments donnerait l'estimateur $\bar{X}$ puisque, par symétrie, μ est nécessairement la moyenne de la loi. En fait on peut montrer que la médiane est un meilleur estimateur, au sens de l'e.q.m., que la moyenne pour cette loi. En particulier le rapport de la variance de la médiane empirique à celui de la moyenne empirique tend vers $2/3$ quand $n \to \infty$. ■

6.7.3 Reparamétrisation et fonctions du paramètre

Une des propriétés séduisantes de l'EMV est qu'aucun nouveau calcul n'est nécessaire en cas de changement de paramètre. On l'appelle la **propriété d'invariance**.

Proposition 6.10 *Soit $\rho = h(\theta)$ une reparamétrisation, alors l'EMV de ρ est $\hat{\rho}^{MV} = h(\hat{\theta}^{MV})$.*

En effet, soient L_θ et L_ρ les vraisemblances respectives de θ et ρ. Comme h est une bijection, l'ensemble des valeurs prises par $L(\theta)$ quand θ décrit Θ est aussi l'ensemble des valeurs prises par ρ quand il décrit son espace paramétrique que nous notons Ω. Posons $\hat{\rho} = h(\hat{\theta}^{MV})$. On a donc :

$$L_\rho(\hat{\rho}) = L_\theta(h^{-1}(\hat{\rho})) = L_\theta(h^{-1}(h(\hat{\theta}^{MV}))) = L_\theta(\hat{\theta}^{MV})$$

qui reste supérieur ou égal à $L_\theta(\theta)$ pour $\theta \in \Theta$ et donc à $L_\rho(\rho)$ pour $\rho \in \Omega$. $\hat{\rho}$ est donc la valeur (unique) où L_ρ atteint son maximum.

Ainsi, par exemple, $\bar{X}$ étant l'EMV du paramètre p de la loi de Bernoulli, $\bar{X}/(1-\bar{X})$ est l'EMV du rapport $p/(1-p)$. Dans la classe exponentielle nous avons mis en évidence (voir proposition 6.7) une reparamétrisation qui admet un estimateur efficace. Celui-ci, par la proposition 6.9, est l'EMV pour le paramètre correspondant. Pour la loi exponentielle $\bar{X}$ est efficace pour $\theta = 1/\lambda$ et est donc nécessairement l'EMV de θ (et $1/\bar{X}$ est celui de λ).

Pour la loi de Pareto avec a connu, comme $\sum_{i=1}^n \ln(\frac{X_i}{a})$ estime de façon efficace $\frac{1}{\theta}$ (voir exercices), on trouve pour EMV de θ : $\left[\frac{1}{n}\sum_{i=1}^n \ln(X_i/a)\right]^{-1}$ (voir exemple 6.19) alors que l'estimateur UMVUE est $\left[\frac{1}{n-1}\sum_{i=1}^n \ln(X_i/a)\right]^{-1}$. Ici encore on constate qu'on a un léger biais mais qu'on reste proche de l'estimateur sans biais optimal.

Selon la propriété d'invariance, l'EMV du couple (μ, σ) de la loi $\mathcal{N}(\mu, \sigma^2)$ est $(\bar{X}, \widetilde{S})$. Signalons en passant que si $(\bar{X}, S^2)$ est l'estimateur UMVUE pour (μ, σ^2), cela n'est pas vrai de $(\bar{X}, S)$ pour (μ, σ) qui est également biaisé (voir exercices). Dans le domaine du contrôle de qualité on utilise souvent $X_{(n)} - X_{(1)}$,

avec un coefficient qui dépend de n et est tabulé, pour estimer σ, ce qui est moins efficace que S mais évidemment plus rapide.

Note 6.4 On convient d'appeler $h(\hat{\theta}^{MV})$ **l'estimateur du maximum de vraisemblance pour la fonction** $h(\theta)$ du paramètre, qu'elle soit bijective ou non. Ainsi, pour la loi de Gauss de paramètre (μ, σ^2), $\widetilde{S}$ est l'EMV de σ et $\bar{X}+1{,}645\,\widetilde{S}$ est l'EMV du quantile d'ordre 0,95 : $\mu+1{,}645\,\sigma$. On donne une légitimité à cette appellation en introduisant la fonction de vraisemblance $L_1(\delta)$ définie sur l'ensemble des valeurs δ atteintes par $h(\theta)$, qui prend, pour un δ donné, la valeur maximale de $L(\theta)$ pour l'ensemble des valeurs θ telles que $h(\theta) = \delta$, i.e. :

$$L_1(\delta) = \sup_{\theta | h(\theta)=\delta} L(\theta).$$

Alors $h(\hat{\theta}^{MV})$ est la valeur qui maximise cette fonction de vraisemblance $L_1(\delta)$ induite par la fonction $h(\theta)$.

6.7.4 Comportement asymptotique de l'EMV

Dans les exemples qui précèdent nous avons pu constater que, sans être totalement optimal, l'EMV restait très proche de l'estimateur UMVUE quand il existait et ceci d'autant plus que n était grand. Ceci se généralise par la propriété essentielle suivante.

Proposition 6.11 *Soit l'échantillon $X_1, X_2, ..., X_n$ issu de la densité (ou fonction de probabilité) $f(x;\theta)$ où $\theta \in \Theta \subseteq \mathbb{R}$, répondant à certaines conditions de régularité qui garantissent notamment l'existence d'un EMV $\hat{\theta}_n^{MV}$ pour tout n. On considère la suite $\{\hat{\theta}_n^{MV}\}$ quand n croît à l'infini. Alors cette suite est telle que :*

$$\sqrt{n}(\hat{\theta}_n^{MV} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{I(\theta)}).$$

Nous admettrons ce résultat qui est en fait une application indirecte de la loi des grands nombres et du théorème central limite. Les conditions de régularité sont celles de la proposition 6.9 complétées d'autres que nous n'expliciterons pas car elles peuvent varier selon le type de démonstration. Le résultat énoncé dans cette proposition implique les propriétés suivantes :

1. $\hat{\theta}_n^{MV}$ est *asymptotiquement sans biais*, i.e. $E_\theta(\hat{\theta}_n^{MV}) \xrightarrow[n\to\infty]{} \theta$.
2. pour n tendant vers l'infini, la variance de $\hat{\theta}_n^{MV}$ se rapproche de $1/(nI(\theta))$. On dit que $\hat{\theta}_n^{MV}$ est *asymptotiquement efficace*.
3. des propriétés 1 et 2 on déduit que $\hat{\theta}_n^{MV}$ converge vers θ en moyenne quadratique (avec un choix adéquat de conditions de régularité on démontre que $\hat{\theta}_n^{MV}$ converge presque sûrement).

4. $\hat{\theta}_n^{MV}$ tend à devenir gaussien quand n s'accroît.

On résume ces propriétés en disant que l'EMV est un estimateur BAN (*Best Asymptotically Normal).*

L'intérêt de ce résultat est double. D'une part il garantit que l'EMV, moyennant des conditions de régularité, soit de très bonne qualité pour les **grands échantillons** (disons $n > 30$), d'autre part il va permettre une approximation de sa distribution d'échantillonnage par une gaussienne, ce qui sera très utile pour établir des intervalles de confiance (voir chapitre 7). La méthode des moments est loin d'offrir les mêmes garanties et c'est pourquoi **la méthode du MV est la méthode de référence**, notamment dans les logiciels statistiques.

Nous attirons l'attention sur le fait que les dites conditions de régularité, si elles sont suffisantes pour que l'EMV soit BAN dans un cadre général, ne sont pas nécessaires. Ainsi pour la loi exponentielle double, vue dans l'exemple 6.22, la médiane empirique se trouve être également BAN bien que la fonction de vraisemblance ne soit pas dérivable partout. Ceci tient au fait que la dérivabilité n'est pas assurée uniquement pour un ensemble discret de points.

Pour finir signalons deux types d'extension de la proposition 6.11 :

1. elle reste valable pour estimer une fonction $h(\theta)$ deux fois dérivable, en substituant $[h'(\theta)]^2 / I(\theta)$ à $1/I(\theta)$
2. elle s'étend à un paramètre à k dimensions : l'EMV $\hat{\theta}_n^{MV}$ est un vecteur aléatoire tel que $\sqrt{n}(\hat{\theta}_n^{MV} - \theta)$ tende en loi vers la *loi de Gauss multivariée* à k dimensions de moyenne nulle et de matrice des variances-covariances égale à $[\mathbb{I}(\theta)]^{-1}$, l'inverse de la matrice d'information. Nous verrons une application de ce type pour le modèle de régression logistique au chapitre 11.

6.8 Les estimateurs bayésiens

Nous abordons ici l'approche bayésienne qui relève d'une philosophie particulière de la statistique. D'une façon générale on qualifie ainsi toute approche qui **confère à tout paramètre inconnu un statut de variable aléatoire** en stipulant pour celui-ci une distribution sur Θ appelée *loi a priori*. Cette loi peut résulter de la connaissance que l'on peut avoir acquise antérieurement sur le phénomène ou être un simple artifice permettant de mener à bien les calculs. En général on tendra à utiliser une loi *a priori* à laquelle le résultat final sera relativement peu sensible (on définit notamment des lois *a priori* dites «non informatives»). L'espace paramétrique étant généralement continu définissons cette loi par une densité, notée $\pi(\theta)$. Pour simplifier on supposera le paramètre de dimension 1, mais l'extension à une dimension quelconque ne présente pas de difficultés.

Dans ce cadre, $f(x;\theta)$ doit être considérée maintenant comme une densité (ou fonction de probabilité) **conditionnelle** pour la v.a. X étudiée, étant donné une valeur fixée du paramètre θ (il serait donc approprié de l'écrire $f(x|\theta)$). En suivant la *formule de Bayes* qui permet de passer de la loi de probabilité d'un événement A sachant B à la probabilité de B sachant A selon :

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} ,$$

on définit la *loi a posteriori* de θ, c'est-à-dire après avoir pris connaissance des réalisations $(x_1, x_2, ..., x_n)$ de l'échantillon $(X_1, X_2, ..., X_n)$. Ci-après le vecteur des réalisations sera noté $\mathbf{x}$ et l'échantillon sera noté $\mathbf{X}$. Par transcription de la formule de Bayes la densité *a posteriori* est[1] :

$$\pi_{\theta|\mathbf{X}=\mathbf{x}}(\theta) = \frac{f(\mathbf{x};\theta)\pi(\theta)}{\int_\Theta f(\mathbf{x};\theta)\pi(\theta)d\theta}$$

où le dénominateur est la densité marginale de $\mathbf{X}$ au point $\mathbf{x} \in \mathbb{R}^n$ pour le $(n+1)$-uplet aléatoire $(\mathbf{X}, \theta)$. Notons que dans cette formule $f(\mathbf{x};\theta)$ peut être aussi bien une densité qu'une fonction de probabilité.

On prend alors comme estimation bayésienne $\hat{\theta}^B$ de θ, **la moyenne de la loi a posteriori.** L'estimateur bayésien s'obtient en appliquant à $\mathbf{X}$ la fonction associant à une valeur de $\mathbf{x}$ quelconque la valeur $\hat{\theta}^B$ correspondante.

Les avantages de cette approche sont multiples du fait que l'on dispose d'une loi pour θ. Entre autres :

- on peut déterminer aisément un intervalle de valeurs plausibles pour θ (voir chapitre 7 sur l'estimation par intervalles),
- on peut estimer θ selon divers critères d'erreur. Le critère des moindres carrés, par exemple, choisit le nombre $\hat{\theta}^B$ minimisant $E[(\theta - \hat{\theta}^B)^2]$, où ici θ est aléatoire, ce qui correspond à la moyenne de la loi *a posteriori* de θ. Nous l'avons privilégié car il est le plus répandu. Mais on pourrait souhaiter minimiser $E(|\theta - \hat{\theta}^B|)$ ce qui débouche sur la médiane de la loi *a posteriori*.
- on peut estimer toute fonction de θ en calculant directement, pour le critère des moindres carrés, l'espérance de $h(\theta)$ sur la loi *a posteriori*, soit :

$$E\left(h(\theta)\right) = \int_\Theta h(\theta)\, \pi_{\theta|\mathbf{X}=\mathbf{x}}(\theta)d\theta.$$

Exemple 6.23 Soit une loi de Bernoulli $\mathcal{B}(p)$. La densité conjointe au point $(x_1, x_2, ..., x_n)$ étant donné $p \in [0\,,1]$ est :

$$\prod_{i=1}^{n} f(x_i;\theta) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = p^s(1-p)^{n-s}$$

[1] Voir à ce propos l'expression d'une densité conditionnelle en fin de section 3.2.

où $s = \sum_{i=1}^{n} x_i$ est le nombre total de succès observé. Supposons que l'on soit dans l'ignorance totale des valeurs préférentielles pour p et prenons une loi *a priori* $\mathcal{U}[0,1]$, donc : $\pi(p) = 1$ pour $p \in [0\,,1]$. La densité *a posteriori* est :

$$\pi_{p|\mathbf{X}=\mathbf{x}}(p) = \frac{p^s(1-p)^{n-s}.1}{\int_0^1 p^s(1-p)^{n-s}.1\,dp} = cp^s(1-p)^{n-s}$$

où c est la constante appropriée pour avoir une densité. Cette densité est celle d'une loi $Beta(s, n-s)$ vue en section 4.2.9.

D'où l'espérance sur cette loi *a posteriori* que nous choisissons comme estimation de p :

$$\hat{p}^B = \frac{\int_0^1 p^{s+1}(1-p)^{n-s}dt}{\int_0^1 p^s(1-p)^{n-s}dt}.$$

En admettant la formule d'intégration :

$$\int_0^1 x^a(1-x)^b dx = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

où $\Gamma(r+1) = r!$ si r est entier, on obtient :

$$\hat{p}^B = \frac{(s+1)!(n-s)!}{(n+2)!}.\frac{(n+1)!}{s!(n-s)!} = \frac{s+1}{n+2}.$$

Cette estimation correspond à l'estimateur étudié dans l'exemple 6.7 pour lequel il a été montré que l'erreur quadratique moyenne était meilleure que celle de l'estimateur UMVUE S_n/n lorsque p se situe autour de $1/2$. ■

On peut démontrer diverses propriétés générales des estimateurs bayésiens. En particulier qu'ils sont convergents quelle que soit la loi *a priori* $\pi(\theta)$ choisie (mais ayant pour support Θ) et même BAN (*best asymptotically normal*) sous des conditions de régularité de la famille $\{f(x;\theta)\}$. On peut également voir sur la formule de la densité *a posteriori*, qu'en raison du théorème de factorisation, cette dernière ne dépendra que d'une statistique exhaustive minimale.

Note 6.5 Étant donné que $\pi(\theta)$ figure au numérateur et au dénominateur de la densité *a posteriori*, il est possible de ne la définir qu'à une constante près. On peut même envisager des fonctions qui ne sont pas des densités (pourvu qu'elles soient positives) considérées alors comme des fonctions de pondération des différentes valeurs possibles de θ. Dans l'exemple ci-dessus on pourrait ainsi prendre la fonction $[p(1-p)]^{-1}$ qui n'est pourtant pas intégrable sur $[0,1]$. Cette fonction donne d'ailleurs s/n comme estimation.

Dans ce chapitre nous avons traité de l'estimation ponctuelle en cherchant à dégager les meilleurs estimateurs. Dans le chapitre suivant on considère des

«fourchettes» d'estimation où la qualité de précision des estimateurs, en particulier la variance pour les estimateurs sans biais, jouera un rôle central.

Pour approfondir l'estimation ponctuelle on pourra consulter l'ouvrage de Lehmann et Casella (1998).

6.9 Exercices

Exercice 6.1 Montrer que la famille de lois $\mathcal{BN}(r,p)$ avec r connu appartient à la classe exponentielle et que $d(x) = x$. Faire de même pour les lois de Poisson, exponentielle et gamma (avec r connu).

Exercice 6.2 Montrer que la famille des lois de Pareto avec a connu appartient à la classe exponentielle et mettre en évidence sa fonction $d(x)$. L'estimateur des moments dépend-il d'une statistique exhaustive minimale ?

Exercice 6.3 Montrer que la famille des lois bêta appartient à la classe exponentielle et mettre en évidence les fonctions $d_1(x)$ et $d_2(x)$. Donner l'estimateur des moments de (α, β). Dépend-il d'une statistique exhaustive minimale ?

Exercice 6.4 Soit la famille $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu. Déduire la loi de S de celle de $(n-1)S^2/\sigma^2$. Montrer que S est un estimateur biaisé de σ et proposer un estimateur sans biais.

Aide : calculer $E(S)$ directement sur la loi de $(n-1)S^2/\sigma^2$.

Exercice 6.5 Soit une famille de lois $\{f(x;\theta)\}$ telle que $f(x;\theta)$ s'écrive $f(x-\theta)$ où f ne dépend pas de θ. (On dit que θ est un paramètre de positionnement et on note qu'on a également $F(x;\theta) = F(x-\theta)$ pour la fonction de répartition). Supposons, de plus, que f soit une fonction paire (lois symétriques par rapport à θ). Les résultats suivants seront établis pour n impair (mais sont valables pour $n = 2k$ en définissant la médiane de façon unique par $\frac{1}{2}(X_{(k)} + X_{(k+1)})$).

1. Établir la fonction de répartition, puis la densité de la médiane empirique d'un n-échantillon.
2. Montrer que sa loi est également symétrique par rapport à θ.
3. En déduire que la médiane empirique est un estimateur sans biais pour θ.

Exercice 6.6 Soit la famille de loi $\mathcal{E}(\lambda)$. Comparer en e.q.m. les estimateurs $\bar{X}$ et $\sum_{i=1}^{n} X_i/(n+1)$ pour estimer $\frac{1}{\lambda}$.

Aide : utiliser les résultats de l'exemple 6.11.

Exercice 6.7 Pour une loi mère quelconque ayant un moment d'ordre 4 comparer les e.q.m. de S^2 et de $\widetilde{S}^2$ et donner une condition pour que le second domine le premier.

Aide : voir la variance de S^2 dans l'exercice 5.4.

Exercice 6.8 Montrer que $X_{(n)}$ est convergent en probabilité pour θ dans la famille $\mathcal{U}[0,\theta]$.

Aide : à partir de la fonction de répartition de $X_{(n)}$ écrire la probabilité de $(|X_{(n)} - \theta| < \epsilon)$.

Exercice 6.9 Montrer pour la famille $\mathcal{E}(\lambda)$ que l'estimateur UMVUE de λ domine (en e.q.m.) l'estimateur des moments. On supposera $n > 2$.

Aide : utiliser les résultats de l'exemple 6.11.

Exercice 6.10 Soit la famille de lois (de Raleigh) de densités :

$$f(x;\theta) = \frac{x}{\theta} \exp\left\{-\frac{x^2}{2\theta}\right\}, \quad x \geq 0, \quad \theta > 0\,.$$

Appartient-elle à la classe exponentielle ? Donner une statistique exhaustive minimale. Calculer son espérance mathématique et en déduire un estimateur sans biais, efficace pour θ.

Exercice 6.11 Soit la famille de Pareto de paramètre a connu et θ inconnu.

1. Montrer que $\ln(\frac{X}{a})$ suit une loi $\mathcal{E}(\theta)$.
2. Montrer que
$$\frac{n-1}{\sum_{i=1}^{n} \ln(X_i/a)}$$
est UMVUE pour θ.
Aide : en calculer l'espérance en s'inspirant du résultat de l'exemple 6.11.
3. Montrer que $\frac{1}{n}\sum_{i=1}^{n} \ln(\frac{X_i}{a})$ est sans biais et efficace pour estimer $1/\theta$.

Exercice 6.12 Soit la famille de lois de Pareto où θ est connu mais a est inconnu.

1. Constater qu'elle n'appartient pas à la classe exponentielle.
2. Donner l'estimateur des moments pour a.
3. Trouver une statistique exhaustive minimale.
4. Identifier la loi de $X_{(1)}$. En déduire un estimateur sans biais fonction de $X_{(1)}$.

Exercice 6.13 Calculer la borne de Cramer-Rao pour la famille $\mathcal{N}(\mu\ ,1)$ où μ est inconnu. Même question pour la famille $\mathcal{N}(0\ ,\sigma^2)$ où σ^2 est inconnu. Dans chaque cas montrer que l'estimateur naturel est efficace.

Exercice 6.14 Soit la famille des lois de Bernoulli $\mathcal{B}(p)$. Donner la borne de Cramer-Rao pour un estimateur sans biais du rapport $p/(1-p)$.

Exercice 6.15 Calculer la borne de Cramer-Rao pour la famille des lois de Cauchy

$$f(x;\theta) = \frac{1}{\pi[1+(x-\theta)^2]}, \quad x \in \mathbb{R}\,.$$

Aide : $\frac{1}{8}\arctan x + \frac{1}{8}\frac{x(x^2-1)}{(x^2+1)^2}$ est primitive de $\frac{x^2}{(1+x^2)^3}$.

Sachant que la médiane empirique M_n est telle que $\sqrt{n}(M_n-\theta)$ converge en loi vers la loi $\mathcal{N}(0\ ,\frac{\pi^2}{4})$ vérifier qu'elle n'est pas un estimateur BAN.

Exercice 6.16 Pour la famille $\mathcal{U}[0\ ,\theta]$ comparer les e.q.m. des 3 estimateurs : moments, EMV et $\frac{n+1}{n}X_{(n)}$ qui est la statistique UMVUE (voir note 6.2).

Exercice 6.17 On considère la loi de Poisson tronquée de la valeur 0. Déterminer sa fonction de probabilité. Trouver l'estimation du MV pour un échantillon de taille 15 de moyenne 3 (ceci nécessite une approximation numérique).

Exercice 6.18 Trouver l'EMV pour le paramètre θ de la loi de Pareto avec a connu. Quel est son biais ? (Aide : voir d'abord l'exercice 6.11). Calculer la borne de Cramer-Rao et constater que l'EMV est asymptotiquement efficace.

Exercice 6.19 Soit la famille de densités

$$f(x;\rho) = \rho(\rho+1)x(1-x)^{\rho-1} \ \text{ pour } \ x \in [0,1]\,,$$

où $\rho > 0$. Donner l'EMV pour ρ.

Exercice 6.20 Soit la famille $\mathcal{BN}(r,p)$ où r est connu. Donner l'EMV pour p sur la base d'une seule observation. Même question pour la famille $B(n,p)$ où n est connu.

Exercice 6.21 (capture-recapture) Un étang contient N poissons où N est inconnu. M (connu) poissons ont été marqués. On pêche (sans remise) jusqu'à ce qu'on obtienne le premier poisson marqué. Soit X le nombre de poissons qu'on doit ainsi pêcher. Donner la loi de X en supposant un tirage aléatoire sans remise. En déduire l'équation de vraisemblance de N associée à une (seule) observation x de X. Application : résoudre numériquement avec $M = 100$ et $x = 3$ pour donner une estimation de N.

Exercice 6.22 Donner l'estimateur du MV pour le paramètre λ d'une loi $\Gamma(r,\lambda)$ où r est connu. Donner une approximation de sa loi pour n grand.

Exercice 6.23 Soit la famille des lois de Bernoulli $\mathcal{B}(p)$. Donner la loi *a posteriori* pour p en utilisant une loi *a priori* proportionnelle à $\sqrt{p(1-p)}$ et en déduire l'estimation bayésienne de p.

Aide : on utilisera la formule d'intégration de l'exemple 6.23 et la relation $\Gamma(a+1) = a\Gamma(a)$. Généraliser à une densité *a priori Beta*(α,β).

Chapitre 7

Estimation paramétrique par intervalle de confiance

7.1 Définitions

Dans le chapitre précédent, l'objectif était de donner une valeur unique pour estimer le paramètre inconnu θ. Dans ce chapitre, nous souhaitons donner un ensemble de valeurs plausibles pour θ essentiellement sous forme d'un intervalle. Dans le vocabulaire courant, pour les sondages notamment, c'est l'idée de «fourchette».

Il y a évidemment un lien entre l'approche ponctuelle et l'approche par intervalle, la seconde s'appuyant pour beaucoup sur les résultats de la première. Si l'on s'en tient aux estimateurs sans biais, un estimateur de variance minimale restera au plus proche de θ et on imagine qu'il sera un bon point de départ pour fournir un encadrement. D'autre part, on s'attend à ce que sa variance soit déterminante pour la largeur de l'intervalle. Cependant nous n'approfondirons pas vraiment la notion d'optimalité pour un intervalle de confiance et consacrerons l'essentiel des développements à la construction de tels intervalles. Pour celle-ci nous verrons tout d'abord une méthode générale exacte, mais qui est subordonnée à l'existence d'une «fonction pivot», et ensuite une méthode asymptotique de portée plus générale, reposant sur une approximation gaussienne en particulier via l'estimateur du maximum de vraisemblance.

Après l'approche générale nous établirons les intervalles de confiance classiques pour les moyennes et variances dans le cas gaussien et pour les proportions dans le cas Bernoulli. La méthode des quantiles sera développée pour indiquer une procédure applicable aux petits échantillons (notamment dans les cas Bernoulli et Poisson).

Par ailleurs, un mode de construction, fondé sur une procédure de test, sera vu ultérieurement dans le chapitre 9 (section 9.8).

Définition 7.1 *Soit $X_1, X_2, \cdots, X_n$ un échantillon aléatoire issu d'une loi de densité (ou fonction de probabilité) $f(x;\theta)$ où $\theta \in \Theta$ est un paramètre inconnu de dimension 1. On appelle* ***procédure d'intervalle de confiance de niveau*** *γ tout couple de statistiques (T_1, T_2) tel que, quel que soit $\theta \in \Theta$, on ait :*

$$P_\theta(T_1 \leq \theta \leq T_2) \geq \gamma\,.$$

En pratique on choisira γ assez élevé : couramment $\gamma = 0{,}95$. Ainsi, il y a une forte probabilité pour que **l'intervalle à bornes aléatoires** $[T_1, T_2]$ contienne la vraie valeur de θ. De façon imagée, on peut dire que dans l'univers des échantillons possibles, pour une proportion au moins γ d'entre eux, on obtient un intervalle qui contient θ.

Dans certaines situations, on peut n'être intéressé qu'à établir une borne inférieure ou une borne supérieure pour θ, T_1 ou T_2 étant rejeté à l'infini. On parle alors d'intervalle de confiance unilatéral (par opposition à « bilatéral »).

Exemple 7.1 Prenons l'exemple quelque peu artificiel d'une loi mère gaussienne dont la variance serait connue et supposons qu'elle soit égale à 1. On a, par centrage-réduction de la moyenne empirique du n-échantillon :

$$\frac{\bar{X} - \mu}{\frac{1}{\sqrt{n}}} \rightsquigarrow \mathcal{N}(0\,;1)$$

d'où :

$$P(-1{,}96 < \frac{\bar{X} - \mu}{\frac{1}{\sqrt{n}}} < 1{,}96) = 0{,}95$$

$$P(-\frac{1{,}96}{\sqrt{n}} < \bar{X} - \mu < \frac{1{,}96}{\sqrt{n}}) = 0{,}95\,.$$

L'événement $(-1{,}96/\sqrt{n} < \bar{X} - \mu)$ est équivalent à $(\mu < \bar{X} + 1{,}96/\sqrt{n})$ et, de même, $(\bar{X} - \mu < 1{,}96/\sqrt{n})$ équivaut à $(\bar{X} - 1{,}96/\sqrt{n} < \mu)$. On voit finalement que l'événement $(-1{,}96/\sqrt{n} < \bar{X} - \mu < 1{,}96/\sqrt{n})$ est identique à l'événement $(\bar{X} - 1{,}96/\sqrt{n} < \mu < \bar{X} + 1{,}96/\sqrt{n})$, d'où :

$$P(\bar{X} - \frac{1{,}96}{\sqrt{n}} < \mu < \bar{X} + \frac{1{,}96}{\sqrt{n}}) = 0{,}95$$

et ceci quel que soit μ, ce qui prouve que $[\bar{X} - 1{,}96/\sqrt{n}\,,\ \bar{X} + 1{,}96/\sqrt{n}]$ constitue une procédure d'intervalle de confiance (IC) de niveau 0,95. On voit sur cet exemple que la «largeur» de l'intervalle est proportionnelle à l'écart-type $1/\sqrt{n}$ de l'estimateur ponctuel $\bar{X}$ pour μ. ■

Note 7.1 Pour les cas continus, comme dans l'exemple précédent, on peut espérer atteindre exactement le niveau γ que l'on s'est fixé, du fait de la continuité des fonctions de répartition. Pour les cas discrets, cependant, un niveau de probabilité donné peut ne pas être atteint en raison des sauts de discontinuité. Nous donnerons plus loin un exemple illustrant cela. On se devra alors d'avoir une attitude conservatrice, c'est-à-dire d'utiliser une procédure garantissant que $[T_1, T_2]$ ait une probabilité de couvrir θ qui soit au moins égale au niveau nominal γ. C'est pourquoi il est nécessaire que γ apparaisse comme une borne minimale de probabilité dans la définition 7.1. Notons encore, au vu de l'exemple 7.1, que le choix de l'intervalle n'est pas unique. On aurait également pu prendre :

$$P(z_{0,03} < \frac{\bar{X} - \mu}{\frac{1}{\sqrt{n}}} < z_{0,98}) = 0,95$$

comme point de départ ou tout autre couple de quantiles $(z_\alpha, z_{0,95+\alpha})$ avec $\alpha \in [0\,;0{,}05]$. L'usage veut, même si la procédure n'est pas nécessairement optimale, que l'on choisisse, comme nous l'avons fait, le couple $(z_{\frac{1-\gamma}{2}}, z_{\frac{1+\gamma}{2}})$. En fait ce choix est celui qui donne la largeur minimale lorsque la densité de la loi utilisée pour les quantiles est symétrique et n'a qu'un seul mode.

Définition 7.2 *Dans le contexte de la définition 7.1, soit $x_1, x_2, \cdots, x_n$ une réalisation de $X_1, X_2, \cdots, X_n$ conduisant à la réalisation (t_1, t_2) de (T_1, T_2). Alors l'intervalle $[t_1, t_2]$ est appelé* ***intervalle de confiance de niveau*** *γ pour θ et l'on note :*

$$IC_\gamma(\theta) = [t_1, t_2]\,.$$

L'intervalle de confiance est donc **l'application numérique de la procédure** suite à la réalisation de l'échantillon. Supposons que dans l'exemple précédent, avec un échantillon de taille 9, on ait observé la valeur 6 pour la moyenne de cet échantillon, alors :

$$IC_{0,95}(\mu) = \left[6 - \frac{1{,}96}{\sqrt{9}}, 6 + \frac{1{,}96}{\sqrt{9}}\right] \simeq [5{,}35\,;6{,}65].$$

On ne peut dire à proprement parler (même si la tentation est forte) que cet IC contient μ avec probabilité 0,95 du fait qu'il s'agit d'une réalisation. Soit il contient μ, soit il ne le contient pas. C'est la procédure choisie en amont qui garantit **a priori** une telle probabilité. C'est pourquoi on parle d'un niveau de confiance et non de probabilité pour un IC.

Note 7.2 Lorsque l'intervalle sera symétrique par rapport à l'estimation ponctuelle on pourra aussi noter, comme pour l'application ci-dessus :

$$IC_{0,95}(\mu) = 6 \pm \frac{1{,}96}{\sqrt{9}}.$$

Note 7.3 On remarquera que le fait d'augmenter le niveau de confiance accroît la largeur de l'intervalle et qu'il n'est pas possible de donner un intervalle certain autre que Θ dans sa totalité.

Note 7.4 S'il s'agit d'estimer une fonction $h(\theta)$ bijective, par exemple strictement croissante, du paramètre θ, il suffit de prendre l'intervalle $[h(t_1), h(t_2)]$.

Nous introduisons maintenant la méthode de la fonction pivot, qui permet de résoudre la plupart des cas classiques.

7.2 Méthode de la fonction pivot

Définition 7.3 *Soit le contexte de la définition 7.1.*
Une fonction $g(X_1, X_2, \cdots, X_n; \theta)$ est appelée ***fonction pivot*** *si :*

1. *la loi de $g(X_1, X_2, \cdots, X_n; \theta)$ est connue et ne dépend pas de θ ,*
2. *pour tous réels u_1 et u_2 tels que $u_1 \leq u_2$ et tout $(x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$, la double inégalité*
$$u_1 \leq g(x_1, x_2, \cdots, x_n; \theta) \leq u_2$$
peut se résoudre (ou «pivoter») en θ selon :
$$t_1(x_1, x_2, \cdots, x_n) \leq \theta \leq t_2(x_1, x_2, \cdots, x_n).$$

Dans l'exemple 7.1, la variable aléatoire $\frac{\bar{X}-\mu}{1/\sqrt{n}}$ était une fonction pivot car pour toute valeur $\bar{x}$ on peut résoudre l'inégalité :

$$u_1 < \frac{\overline{x} - \mu}{1/\sqrt{n}} < u_2$$

en :

$$\bar{x} - \frac{u_2}{\sqrt{n}} < \mu < \bar{x} - \frac{u_1}{\sqrt{n}}\,.$$

Notons que dans cette définition, on peut évidemment se restreindre aux valeurs $(x_1, x_2, \cdots, x_n)$ appartenant à l'ensemble des réalisations possibles pour θ quelconque. Remarquons aussi qu'une fonction pivot n'est pas une statistique car elle contient le paramètre inconnu θ.

Proposition 7.1 *L'existence d'une fonction pivot assure une procédure d'intervalle de confiance de niveau donné quelconque.*

En effet, il suffit de choisir, sur la loi connue, des quantiles u_1 et u_2 tels que :

$$P(u_1 < g(X_1, X_2, \cdots, X_n; \theta) < u_2) \geq \gamma$$

puis de faire « pivoter », pour encadrer θ. C'est ce qui a été effectué dans l'exemple 7.1. Donnons un autre exemple.

Exemple 7.2 Soit $X_1, X_2, \cdots, X_n$ un échantillon de loi mère $\mathcal{E}(\lambda)$. Nous avons vu en section 4.2.3 que $T = \sum_{i=1}^{n} X_i$ suit une loi $\Gamma(n, \lambda)$ et il n'est pas inutile de rappeler que cette statistique est exhaustive minimale. Sa densité est :

$$f_T(t) = \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} \quad (si\ t \geq 0).$$

Considérons la variable aléatoire λT et déterminons sa fonction de densité $f_{\lambda T}$. Avec des notations évidentes on a :

$$F_{\lambda T}(t) = P(\lambda T \leq t) = P(T \leq \frac{t}{\lambda}) = F_T(\frac{t}{\lambda}).$$

D'où, par dérivation :

$$f_{\lambda T}(t) = \frac{1}{\lambda} f_T(\frac{t}{\lambda}) = \frac{1}{(n-1)!} t^{n-1} e^{-t} \quad (si\ t \geq 0)$$

qui est la densité de la loi $\Gamma(n, 1)$ et ne dépend pas de λ. De toute évidence, la double inégalité $u_1 \leq \lambda T \leq u_2$ peut «pivoter» pour isoler le paramètre λ selon $\frac{u_1}{T} \leq \lambda \leq \frac{u_2}{T}$. Pour obtenir une procédure d'intervalle de confiance de niveau, disons, 0,95 il suffit de choisir pour u_1 et u_2 respectivement, les quantiles d'ordre 0,025 et 0,975 de la loi $\Gamma(n, 1)$, c'est-à-dire les valeurs u_1 et u_2 telles que :

$$\int_0^{u_1} \frac{1}{(n-1)!} u^{n-1} e^{-u} du = 0,05 \quad et \quad \int_0^{u_2} \frac{1}{(n-1)!} u^{n-1} e^{-u} du = 0,975$$

que l'on doit lire dans les tables de la fonction de répartition des lois gamma ou déterminer via une fonction ad hoc d'un logiciel statistique (souvent appelée fonction gamma inverse). On a alors :

$$IC_{0,95}(\lambda) = \left[\frac{u_1}{\sum_{i=1}^{n} X_i}, \frac{u_2}{\sum_{i=1}^{n} X_i} \right].$$

■

Nous verrons d'autres illustrations de fonctions pivots dans la section 7.4 qui concerne les IC classiques.

En dehors de ces cas classiques, il n'existera généralement pas de fonction pivot et il est nécessaire d'avoir une procédure de type universel pour couvrir les situations les plus variées, et même complexes. La méthode qui a la portée la plus générale est la méthode asymptotique qui fournit une approximation d'IC d'une façon que nous allons préciser.

7.3 Méthode asymptotique

Plaçons-nous dans le cas le plus général et supposons qu'il existe un estimateur T_n de θ tel que :

$$\frac{T_n - \theta}{s_n(\theta)} \quad \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \quad \mathcal{N}(0\,;1)$$

où $s_n(\theta)$ est une fonction appropriée de θ : le plus souvent l'écart-type de T_n ou une fonction équivalente quand $n \to \infty$. Si la fonction $\frac{T_n-\theta}{s_n(\theta)}$ pivote pour isoler θ on a la procédure d'IC approchée recherchée. Sinon, T_n étant convergente pour θ, moyennant la continuité de la fonction s_n (évidemment quel que soit n), $\frac{T_n-\theta}{s_n(T_n)}$ convergera aussi en loi vers la loi normale centrée-réduite. Alors le pivotement est immédiat pour donner l'IC approximatif :

$$IC_\gamma(\theta) \simeq [t_n - z_{\frac{1+\gamma}{2}}\, s_n(t_n)\,;\, t_n + z_{\frac{1+\gamma}{2}}\, s_n(t_n)]$$

où t_n est la réalisation de T_n.

Cet intervalle est approximatif dans le sens où la procédure correspondante ne garantit pas exactement le niveau γ quel que soit θ pour n fini. Bien qu'il soit difficile de donner un seuil pour n à partir duquel on sera suffisamment proche du niveau γ (disons à 10^{-2} près), on se référera à la règle $n \geq 30$ indiquée pour le théorème central limite. En effet il est clair que c'est ce théorème qui est susceptible de nous fournir un estimateur approprié comme dans l'exemple ci-après.

Exemple 7.3 (IC pour λ de $\mathcal{P}(\lambda)$) Nous avons vu (section 6.6.3) que $\bar{X}_n$, la moyenne de l'échantillon, est un estimateur efficace pour λ du fait que pour cette famille de lois on a $d(x) = x$. Cette statistique est donc particulièrement intéressante pour envisager un IC pour λ. Comme $\sqrt{\lambda}$ est l'écart-type de la loi, le théorème central limite indique que :

$$\frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \quad \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \quad \mathcal{N}(0\,;1)\,.$$

Choisissant un niveau de confiance γ, on a :

$$P(-z_{\frac{1+\gamma}{2}} \leq \frac{\bar{X}_n - \lambda}{\sqrt{\frac{\lambda}{n}}} \leq z_{\frac{1+\gamma}{2}}) \simeq \gamma.$$

La double inégalité se ramène à une inégalité du second degré en λ que l'on peut résoudre :

$$\frac{(\bar{X}_n - \lambda)^2}{\frac{\lambda}{n}} \quad \leq \quad z^2_{\frac{1+\gamma}{2}}$$

$$\text{ou} \quad \lambda^2 - \lambda \left(2\bar{X}_n + \frac{z^2_{\frac{1+\gamma}{2}}}{n}\right) + \bar{X}_n^2 \leq 0\,.$$

Or :

$$\Delta = \left(2\bar{X}_n + \frac{z^2_{\frac{1+\gamma}{2}}}{n}\right)^2 - 4\bar{X}_n^2 = 4\frac{\bar{X}_n z^2_{\frac{1+\gamma}{2}}}{n} + \frac{z^4_{\frac{1+\gamma}{2}}}{n^2} > 0$$

et le polynôme du second degré en λ est négatif entre les racines, d'où la procédure d'IC approximatif :

$$P\left(\bar{X}_n + \frac{z^2_{\frac{1+\gamma}{2}}}{2n} - \sqrt{\frac{\bar{X}_n z^2_{\frac{1+\gamma}{2}}}{n} + \frac{z^4_{\frac{1+\gamma}{2}}}{4n^2}} < \lambda < \bar{X}_n + \frac{z^2_{\frac{1+\gamma}{2}}}{2n} + \sqrt{\frac{\bar{X}_n z^2_{\frac{1+\gamma}{2}}}{n} + \frac{z^4_{\frac{1+\gamma}{2}}}{4n^2}}\right) \simeq \gamma\,.$$

En négligeant sous la racine le terme en $\frac{1}{n^2}$ par rapport à celui en $\frac{1}{n}$, puis celui en $\frac{1}{n}$ par rapport à celui en $\frac{1}{\sqrt{n}}$ on obtient finalement :

$$P\left(\bar{X}_n - z_{\frac{1+\gamma}{2}}\sqrt{\frac{\bar{X}_n}{n}} < \lambda < \bar{X}_n + z_{\frac{1+\gamma}{2}}\sqrt{\frac{\bar{X}_n}{n}}\right) \simeq \gamma.$$

Cet intervalle est précisément celui que l'on aurait obtenu en substituant à λ l'estimateur $\bar{X}_n$ dans l'expression de la variance $\frac{\lambda}{n}$ (conformément à la substitution $s_n(T_n)$ pour $s_n(\theta)$ évoquée plus haut). On voit donc que cette substitution est une approximation du second ordre par rapport au résultat du théorème central limite. Généralement, il en sera ainsi et nous verrons un cas similaire pour l'IC classique sur le paramètre p d'une loi de Bernoulli (voir section 7.4.5). Pour conclure, retenons la formule suivante pour le paramètre λ de la loi de Poisson :

$$IC_{0,95}(\lambda) = \bar{x} \pm 1{,}96\sqrt{\frac{\bar{x}}{n}}$$

qui, en règle pratique, donne une approximation satisfaisante dès que l'on a $\sum_{i=1}^n x_i > 30$. Pour $\sum_{i=1}^n x_i$ plus petit on applique la procédure qui sera développée au cours de l'exemple 7.6 via la méthode des quantiles. ■

La question qui semble se poser pour la mise en œuvre de la méthode asymptotique est celle de l'existence d'un estimateur du type de T_n. En fait, **l'estimateur du maximum de vraisemblance**, moyennant certaines conditions de régularité, fera l'affaire. Nous avons même vu (voir proposition 6.11) qu'il est un estimateur BAN (*best asymptotically normal*) et l'on peut montrer qu'il fournira des IC qui auront une certaine optimalité asymptotique, notamment en termes de largeur d'intervalle. Dans les notations de la section 6.7, nous avons (proposition 6.11) :

$$\frac{\hat{\theta}_n^{MV} - \theta}{\sqrt{\frac{1}{nI(\theta)}}} \quad \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \quad \mathcal{N}(0\,;1).$$

Mis à part quelques cas simples, l'expression $I(\theta)$ de l'information de Fisher sera telle qu'elle ne permettra pas le pivotement. On lui substituera donc $I(\hat{\theta}_n^{MV})$ pour obtenir finalement la formule générale suivante (où $\hat{\theta}^{MV}$ désigne cette fois l'estimation du MV) :

$$IC_\gamma(\theta) \simeq \left[\hat{\theta}^{MV} - \frac{z_{\frac{1+\gamma}{2}}}{\sqrt{nI(\hat{\theta}^{MV})}} , \hat{\theta}^{MV} + \frac{z_{\frac{1+\gamma}{2}}}{\sqrt{nI(\hat{\theta}^{MV})}} \right].$$

Exemple 7.4 (Paramètre θ de la loi de Pareto avec a connu)
Il a été indiqué (section 6.7.3) que l'EMV pour θ est $\hat{\theta}^{MV} = [\frac{1}{n}\sum_{i=1}^{n} \ln(\frac{X_i}{a})]^{-1}$. Par ailleurs, on peut montrer (voir exercices du chapitre 6) que la borne de Cramer-Rao est $\frac{\theta^2}{n}$. On a donc :

$$\frac{\hat{\theta}^{MV} - \theta}{\frac{\theta}{\sqrt{n}}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1).$$

Cette expression permet le pivotement sans qu'il soit nécessaire de recourir à la substitution de $I(\hat{\theta}^{MV})$ pour $I(\theta)$. En effet (en prenant $\gamma = 0{,}95$ pour simplifier) :

$$P\left(-1{,}96 < \frac{\hat{\theta}^{MV} - \theta}{\frac{\theta}{\sqrt{n}}} < 1{,}96 \right) \simeq 0{,}95$$

$$\Leftrightarrow P\left(\frac{\hat{\theta}^{MV}}{1 + \frac{1{,}96}{\sqrt{n}}} < \theta < \frac{\hat{\theta}^{MV}}{1 - \frac{1{,}96}{\sqrt{n}}} \right) \simeq 0{,}95.$$

En utilisant l'approximation $[1 + \frac{1{,}96}{\sqrt{n}}]^{-1} \simeq 1 - \frac{1{,}96}{\sqrt{n}}$ qui néglige le terme en $\frac{1}{n}$ par rapport au terme en $\frac{1}{\sqrt{n}}$, et en faisant de même pour $[1 - \frac{1{,}96}{\sqrt{n}}]^{-1}$, on a :

$$P\left(\hat{\theta}^{MV}(1 - \frac{1,96}{\sqrt{n}}) < \theta < \hat{\theta}^{MV}(1 + \frac{1,96}{\sqrt{n}}) \right) \simeq 0,95.$$

Une fois encore, cette expression est celle que l'on aurait obtenue en substituant d'emblée $\frac{\hat{\theta}^{MV}}{n}$ à $\frac{\theta}{n}$ pour la variance asymptotique de $\hat{\theta}^{MV}$. ■

La construction d'un IC à partir de l'EMV se heurte dans les situations non standard à une difficulté pratique, à savoir la détermination de l'information de Fisher $I(\theta)$. Nous indiquons la façon dont les logiciels statistiques qui fournissent des IC sur les paramètres des divers modèles qu'ils proposent, contournent ce problème y compris lorsque les observations ne sont pas i.i.d.

(par exemple : modèles de régression, modèles de séries chronologiques), la seule exigence étant de connaître la forme de la densité conjointe $f(x_1, x_2, \cdots, x_n; \theta)$. Nous verrons une situation de ce type en régression logistique au chapitre 11.

Résolution numérique pour l'IC fondé sur l'EMV

Reprenons l'expression de $I(\theta)$ donnée en section 6.6.3. Sous certaines conditions de régularité pour la densité $f(x;\theta)$ on a :

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

Il est clair que le calcul d'une telle espérance mathématique peut être inextricable. Toutefois, en vertu de la loi des grands nombres, la v.a.

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(X_i;\theta)$$

converge en probabilité (voire même presque sûrement quand la variance de $\frac{\partial^2}{\partial \theta^2} \ln f(X;\theta)$ existe) vers $I(\theta)$ quand $n \to \infty$. Ceci reste en fait valable en remplaçant θ par $\hat{\theta}_n^{MV}$ et on estimera donc $I(\theta)$ par l'expression ci-dessus calculée au point $\theta = \hat{\theta}_n^{MV}$ et pour les réalisations x_i des X_i, soit :

$$-\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f(x_i;\hat{\theta}_n^{MV}) = -\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}_n^{MV}),$$

où $\ln L(\hat{\theta}_n^{MV})$ est la log-vraisemblance de θ en $\hat{\theta}_n^{MV}$. Numériquement, ceci peut être accompli de façon précise sans calculer explicitement la dérivée seconde de la log-vraisemblance, en donnant de très faibles variations à θ autour de $\hat{\theta}_n^{MV}$.

Le principe de calcul approché se généralise aisément à un paramètre de dimension $k > 1$ à partir de la matrice d'information de Fisher. L'élément (i, j) de cette matrice est estimé par :

$$-\frac{1}{n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln L(\hat{\theta}_n^{MV}),$$

où θ_i et θ_j sont, respectivement, la i-ème et la j-ème composante du vecteur des paramètres. On obtient des IC sur chacune des composantes isolément en considérant les dérivées secondes par rapport à chaque composante. On verra plus loin la notion de région de confiance dans $\mathbb{R}^k$ pour une prise en compte simultanée de toutes les composantes.

Nous abordons maintenant la construction d'IC dans les situations les plus courantes. Elle reposera soit sur la méthode du pivot, soit sur l'approche asymptotique.

7.4 Construction des IC classiques

Sous le terme de « classique » nous présentons les cas de la moyenne et de la variance d'une loi mère gaussienne et le cas du paramètre p d'une loi de Bernoulli. Nous verrons que le résultat pour la moyenne d'une gaussienne peut servir d'approximation pour une loi mère quelconque. Quant au cas de Bernoulli il est d'une grande importance pratique puisqu'il traite des « fourchettes » d'estimation de proportions dans les sondages. Nous rencontrerons également des situations nouvelles de comparaisons entre deux lois (ou, en pratique, deux populations) distinctes. Les résultats qui suivent exploitent ceux établis au chapitre 5 sur les distributions d'échantillonnage. Pour simplifier les écritures, nous prendrons, comme c'est l'usage, des IC de niveau $\gamma = 0{,}95$ faisant donc intervenir les quantiles d'ordres 0,025 et 0,975 , le passage à une autre valeur de γ étant évident.

Nous proposons dans la section des exercices quelques « exercices appliqués » permettant d'illustrer l'intérêt des intervalles de confiance.

7.4.1 IC pour la moyenne d'une loi $\mathcal{N}(\mu, \sigma^2)$

Nous abordons d'emblée le cas où (μ, σ^2) est un paramètre de dimension 2 inconnu, mais nous nous intéresserons ici uniquement à un encadrement pour μ indépendamment de σ^2. Nous reviendrons ensuite brièvement sur le cas plus simple, mais peu réaliste, où σ^2 est supposé connu. Rappelons le résultat du théorème 5.2 :

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \underset{approx}{\rightsquigarrow} t(n-1).$$

Cette v.a. est de toute évidence une fonction pivot pour μ et nous obtenons un IC comme suit, les développements étant de même nature que dans l'exemple 7.1 :

$$P\left(-t_{0,975}^{(n-1)} < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < t_{0,975}^{(n-1)}\right) = 0{,}95$$

où $t_\alpha^{(n-1)}$ est la notation adoptée pour le quantile d'ordre α de la loi de Student à $n-1$ d.d.l. (degrés de liberté) dont nous rappelons que, comme la loi de Gauss, elle est symétrique par rapport à 0. Il s'ensuit que :

$$P\left(\bar{X} - t_{0,975}^{(n-1)} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{0,975}^{(n-1)} \frac{S}{\sqrt{n}}\right) = 0{,}95$$

et le **résultat très classique** :

$$IC_{0,95}(\mu) = \left[\bar{x} - t_{0,975}^{(n-1)} \frac{s}{\sqrt{n}} \ , \ \bar{x} + t_{0,975}^{(n-1)} \frac{s}{\sqrt{n}}\right].$$

Notons que les quantiles des lois de Student sont donnés dans toutes les tables statistiques usuelles. La largeur de cet IC, dont on peut montrer qu'elle est minimale par rapport à d'autres éventuelles procédures, dépend d'une part de la taille d'échantillon et d'autre part de la dispersion même de la loi mère (ou, en pratique, de la variable étudiée dans la population) à travers l'estimation s de son écart-type σ. Plus la population est homogène et plus la taille d'échantillon est élevée, plus l'estimation sera précise.

Les praticiens utilisent cette formule sans se soucier de la « normalité » de la loi mère. Ceci est, de fait, justifié d'une part grâce au théorème central limite qui assure, avec des conditions généralement réalistes, que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ soit à peu près gaussien dès lors que n est assez grand (rappelons que $n > 30$ est, en pratique, bien suffisant) et, d'autre part, grâce à la convergence de la variance d'échantillon S_n^2 en tant qu'estimateur de σ^2. En réalité le deuxième point n'est pas si clair. En effet, si cette convergence est opérante pour de grands échantillons, disons plus d'une centaine d'observations (auquel cas on applique simplement une approximation par la loi de Gauss), on peut se poser la question pour de plus petits échantillons dans la mesure où, pour une loi mère quelconque, $\frac{(n-1)S_n^2}{\sigma^2}$ peut suivre une loi qui s'écarte sensiblement de la loi $\chi^2(n-1)$ théoriquement requise (voir théorème 5.1). Néanmoins on a montré que l'approximation de Student reste relativement satisfaisante.

Ceci explique le caractère quasi universel de la formule pour $IC_{0,95}(\mu)$, dont on peut dire qu'elle fournit un IC approché dans des situations non paramétriques puisqu'elle est valable pour une grande variété de lois (pour le théorème central limite, il suffit que la variance existe, ce qui est aussi suffisant pour la convergence de S_n^2, selon la proposition 6.1). Cependant, dans les cas paramétriques simples, les formules d'IC établies en tenant compte des spécificités de la famille considérée seront plus précises. Ainsi, s'il s'agit d'estimer la moyenne λ d'une loi de Poisson, le résultat obtenu dans l'exemple 7.3 est préférable car il intègre le fait que la variance de la loi est λ et qu'elle n'a pas besoin d'être estimée indépendamment par la variance de l'échantillon.

Revenons maintenant sur la situation où σ^2 **est connu** qui, bien que présentée dans tous les ouvrages, est un cas d'école car rares sont les situations pratiques de ce type. Elles ne sont toutefois pas inexistantes. Ainsi certaines machines-outils devant usiner des pièces selon une certaine cote provoquent, lorsqu'elles se dérèglent, un déplacement de la valeur moyenne mais conservent le même aléa, c'est-à-dire la même variance.

En fait, le cas où σ^2 est connu a été traité dans l'exemple 7.1 où, par commodité, on a supposé $\sigma^2 = 1$. L'IC obtenu est donc :

$$IC_{0,95}(\mu) = \left[\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}}\,, \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}\right].$$

Par contraste avec le précédent IC où σ^2 est inconnu, les quantiles sont à lire sur la loi de Gauss du fait que σ^2 n'a pas à être estimé.

7.4.2 IC pour la variance σ^2 d'une loi de Gauss

Nous supposerons que μ est également inconnu. Le cas d'école où μ est connu est proposé en exercice et s'opère par une voie analogue. Reprenons le résultat du théorème 5.1 :

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1),$$

d'où :

$$P\left(\chi^{2\ (n-1)}_{0,025} < \frac{(n-1)S^2}{\sigma^2} < \chi^{2\ (n-1)}_{0,975}\right) = 0{,}95$$

où $\chi^{2\ (n-1)}_{\alpha}$ dénote le quantile d'ordre α de la loi $\chi^2(n-1)$. Ces quantiles se trouvent dans les tables statistiques ordinaires. On peut directement isoler σ^2 pour obtenir :

$$P\left(\frac{(n-1)S^2}{\chi^{2\ (n-1)}_{0,975}} < \sigma^2 < \frac{(n-1)S^2}{\chi^{2\ (n-1)}_{0,025}}\right) = 0{,}95$$

et

$$IC_{0,95}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^{2\ (n-1)}_{0,975}}\,,\ \frac{(n-1)s^2}{\chi^{2\ (n-1)}_{0,025}}\right].$$

Cet intervalle de confiance est **peu robuste** vis-à-vis de l'hypothèse gaussienne, contrairement à celui sur μ. On ne peut donc l'utiliser dans des situations où la loi mère diffère d'une loi normale. Ceci est vrai même pour une grande taille d'échantillon car on montre (voir section 8.2.2) que la loi asymptotique de S^2 (plus précisément de $\sqrt{n}(S^2-\sigma^2)$) dépend de la loi mère.

Note 7.5 Suivant la note 7.4, on peut déduire un IC pour l'**écart-type** σ de celui sur la variance :

$$IC_{0,95}(\sigma) = \left[\frac{\sqrt{(n-1)}s}{\sqrt{\chi^{2\ (n-1)}_{0,975}}}\,,\ \frac{\sqrt{(n-1)}s}{\sqrt{\chi^{2\ (n-1)}_{0,025}}}\right].$$

Au passage, on peut comparer la variabilité de l'écart-type empirique S et de la moyenne empirique $\bar{X}$, pour une loi de Gauss tout du moins. En première approximation, en appliquant la formule pour une fonction d'une v.a. établie en section 2.6, $V(S) \simeq \frac{1}{4\sigma^2}V(S^2) = \frac{\sigma^2}{2(n-1)}$ alors que $V(\bar{X}) = \frac{\sigma^2}{n}$. La fluctuation de S est plus faible, ce qui se retrouve au niveau des précisions des IC (voir exercices).

7.4.3 IC sur la différence des moyennes de deux lois de Gauss

Nous considérons ici deux lois mères (en pratique, souvent, deux populations) et souhaitons construire un IC sur la différence de leurs moyennes. Citons comme exemples l'écart entre la taille moyenne des filles et des garçons à l'âge de douze ans, l'écart de revenu moyen des actifs entre telle et telle région. Pour cela on dispose de **deux échantillons aléatoires indépendants** pris dans chaque population (le fait de prendre une sœur et un frère pour l'exemple de la taille ne respecterait pas cette hypothèse d'indépendance des deux échantillons).

La procédure classique que nous allons développer suppose que **les deux lois ont même variance** σ^2. Soit un échantillon de taille n_1 issu de la loi $\mathcal{N}(\mu_1, \sigma^2)$ et un échantillon, indépendant du premier, de taille n_2, issu de la loi $\mathcal{N}(\mu_2, \sigma^2)$. Soit $\bar{X}_1$ et S_1^2, moyenne et variance empiriques du premier échantillon et de même $\bar{X}_2$ et S_2^2 pour le deuxième échantillon. On a :

$$\bar{X}_1 \rightsquigarrow \mathcal{N}(\mu_1, \frac{\sigma^2}{n_1})$$
$$\bar{X}_2 \rightsquigarrow \mathcal{N}(\mu_2, \frac{\sigma^2}{n_2})$$

et

$$\bar{X}_1 - \bar{X}_2 \rightsquigarrow \mathcal{N}(\mu_1 - \mu_2\,,\, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$$

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow \mathcal{N}(0\,;1)\,.$$

Le problème qui se pose est celui de l'estimation de σ que l'on effectue, en fait, via σ^2. Sachant que $\frac{(n_1-1)S_1^2}{\sigma^2} \rightsquigarrow \chi^2(n_1-1)$ et $\frac{(n_2-1)S_2^2}{\sigma^2} \rightsquigarrow \chi^2(n_2-1)$, l'indépendance des deux échantillons entraîne (voir proposition 5.7) que :

$$\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \rightsquigarrow \chi^2(n_1+n_2-2)\,.$$

En faisant le rapport de la v.a. $\bar{X}_1 - \bar{X}_2$ centrée-réduite à la racine carrée de la v.a. ci-dessus divisée par ses degrés de liberté, et en posant :

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\,,$$

on obtient :

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t(n_1+n_2-2)$$

(voir les développements similaires du théorème 5.2). La fonction ci-dessus est une fonction pivot qui aboutit immédiatement à :

$$IC_{0,95}(\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm t_{0,975}^{(n_1+n_2-2)}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

où :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

est la variance empirique pondérée en fonction des tailles d'échantillon respectives.

Qu'en est-il de la condition très restrictive d'égalité des variances ? En fait, on a pu montrer que celle-ci n'est pas si cruciale si les tailles d'échantillons n_1 et n_2 diffèrent peu. Dans ce cas un facteur 2 pour le rapport des variances reste acceptable. En revanche si n_1 et n_2 diffèrent substantiellement la formule ci-dessus s'applique mal quand les variances ne sont pas proches. Alors on peut effectuer les mêmes développements que précédemment en introduisant les variances respectives des deux lois σ_1^2 et σ_2^2 pour obtenir :

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \rightsquigarrow \mathcal{N}(0\,;1)$$

et, si les tailles d'échantillons sont élevées, disons au-delà d'une centaine, conserver une approximation raisonnable en substituant à σ_1^2 et σ_2^2 leurs estimations s_1^2 et s_2^2, d'où :

$$IC_{0,95}(\mu_1 - \mu_2) \simeq (\bar{x}_1 - \bar{x}_2) \pm 1{,}96 \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

On remarquera que si $n_1 = n_2$ cette formule est identique à celle du cas où $\sigma_1^2 = \sigma_2^2$ (mis à part les quantiles d'ordre 0,975 qui seront cependant pratiquement identiques pour les grands échantillons). D'autres formules d'approximation plus précises ont été développées, mais elles donnent des résultats numériques proches de ceux obtenus avec l'hypothèse d'égalité des variances ce qui encourage peu leur utilisation par les praticiens.

Indiquons qu'il existe un usage assez répandu consistant à effectuer au préalable un test de l'hypothèse d'égalité des variances comme proposé en section 9.7.4. Même si l'on peut admettre que cela a l'avantage de constituer un garde-fou, cette procédure ne fournit pas une garantie suffisante quant à l'applicabilité de la formule classique en cas d'acceptation de l'hypothèse par le test.

Quant à l'usage de celle-ci en dehors des conditions de «normalité» des deux lois, il est acceptable pour les mêmes raisons que celles exposées dans le cas

d'une seule loi (section 7.4.1). En résumé, le point critique est une différence trop sensible des dispersions des deux lois.

Cas des échantillons appariés

Dans la mesure où cela est possible, on gagnera en précision (i.e. en largeur d'intervalle) en associant les deux échantillons par paires ayant les mêmes valeurs sur une ou plusieurs variables auxiliaires, dites *variables de contrôle.* Le gain sera d'autant plus important que ces variables auxiliaires seront liées à la variable étudiée. S'il s'agit, par exemple, de comparer les effets de deux molécules sur la réduction de l'hypertension on mettra en œuvre un plan d'expérience associant des paires d'individus de même âge, même sexe, même niveau d'hypertension initial. Souvent il s'agit de *mesures répétées* sur le même échantillon, l'appariement étant alors parfait.

Les développements précédents ne sont plus possibles du fait que les deux échantillons ne sont plus indépendants. On contourne ce problème en raisonnant sur la v.a. D «différence entre individus appariés» pour se ramener au cas d'une seule loi.

Ceci est justifié du fait que $E(D) = E(X_1 - X_2) = \mu_1 - \mu_2$. En notant $\bar{d}$ la moyenne des différences entre les n paires observées et s_d l'écart-type de ces différences, on a :

$$IC_{0,95}(\mu_1 - \mu_2) = \bar{d} \pm t_{0,975}^{(n-1)} \frac{s_d}{\sqrt{n}}.$$

Les considérations de «robustesse» de la formule par rapport à l'hypothèse de normalité vues en section 7.4.1 restent valables.

7.4.4 IC sur le rapport des variances de deux lois de Gauss

On considère le rapport $\frac{\sigma_1^2}{\sigma_2^2}$ pour les lois $\mathcal{N}(\mu_1, \sigma_1)$ et $\mathcal{N}(\mu_2, \sigma_2)$. Comme :

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi^2(n_1 - 1) \quad \text{et} \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi^2(n_2 - 1)\,,$$

on a, par application des résultats de la section 5.5 :

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \quad \rightsquigarrow \quad F(n_1 - 1, n_2 - 1).$$

D'où :

$$P\left(F_{0,025}^{(n_1-1,n_2-1)} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{0,975}^{(n_1-1,n_2-1)}\right) = 0{,}95\,,$$

soit finalement après pivotement et compte tenu du fait (voir proposition 5.11) que $F_{\alpha}^{(\nu_1,\nu_2)} = 1/F_{1-\alpha}^{(\nu_2,\nu_1)}$,

$$IC_{0,95}(\frac{\sigma_1^2}{\sigma_2^2}) = \left[\frac{s_1^2}{s_2^2}F_{0,025}^{(n_2-1,n_1-1)}\,,\ \frac{s_1^2}{s_2^2}F_{0,975}^{(n_2-1,n_1-1)}\right].$$

Comme pour la procédure relative à une variance (section 7.4.2), cette formule n'est pas robuste lorsque les lois s'écartent de lois gaussiennes. Son usage est donc très limité.

7.4.5 IC sur le paramètre p d'une loi de Bernoulli

Rappelons, pour les applications, que p peut être la probabilité à estimer pour l'occurrence d'un événement (succès). Dans le cas d'un sondage aléatoire simple sans remise dans une grande population (taux de sondage inférieur à 0,10, voir section 3.7), p est la proportion d'individus possédant un certain caractère dans cette population. C'est pourquoi on parle généralement d'**intervalle de confiance pour une proportion.** Les enquêtes estimant pour l'essentiel des proportions (ou pourcentages), les résultats qui suivent vont fournir les «fourchettes» de précision des sondages.

La statistique sur laquelle se fondent les résultats est S_n le nombre total de succès parmi les n répétitions dont la loi exacte est $\mathcal{B}(n,p)$. C'est une statistique exhaustive minimale pour le paramètre p de la loi mère de Bernoulli. Comme c'est le plus souvent le cas pour des distributions discrètes, on ne peut mettre en évidence une fonction pivot. Pour obtenir des IC exacts on doit recourir à la méthode des quantiles exposée en section 7.5. Cette méthode préside à l'élaboration d'abaques et de tables ainsi qu'aux résultats fournis par les logiciels.

Pour l'heure nous présentons le résultat classique obtenu par l'approche asymptotique qui s'applique dans la plupart des cas du fait des tailles d'échantillons courantes. Suite au théorème central limite nous avons vu (section 5.8.3) que la loi $\mathcal{B}(n,p)$ de S_n peut être approchée convenablement par la loi de Gauss $\mathcal{N}(np, np(1-p))$ pourvu que $np > 5$ et $n(1-p) > 5$. D'où :

$$\frac{S_n - np}{\sqrt{np(1-p)}} \quad \underset{approx}{\rightsquigarrow} \quad \mathcal{N}(0\,;1)$$

qui met en évidence une v.a. de loi asymptotiquement indépendante de p et ainsi, en négligeant la correction de continuité (voir section 5.8.3) :

$$P\left(-1{,}96 < \frac{S_n - np}{\sqrt{np(1-p)}} < 1{,}96\right) \simeq 0{,}95.$$

Pour ce qui concerne le pivotement nous sommes dans une situation analogue à celle rencontrée pour le paramètre λ de la loi de Poisson de l'exemple 7.3,

à savoir que la double inégalité se ramène à une inégalité du second degré en p que l'on peut résoudre :

$$\frac{(S_n - np)^2}{np(1-p)} \leq (1{,}96)^2.$$

Donnons la solution finale pour l'IC, en fonction de la fréquence relative observée $\hat{p}$, réalisation de $\widehat{P}_n = S_n/n$:

$$\frac{\hat{p} + \frac{(1{,}96)^2}{2n}}{1 + \frac{(1{,}96)^2}{n}} \pm \frac{1}{1 + \frac{(1{,}96)^2}{n}} \sqrt{\frac{(1{,}96)^2}{n}\hat{p}(1-\hat{p}) + \frac{(1{,}96)^4}{4n^2}}$$

qui, bien que développée dans divers ouvrages, n'est jamais utilisée.

En effet, les conditions de validité de l'approximation gaussienne $np > 5$ et $n(1-p) > 5$ n'étant pas vérifiables puisque p est inconnu, on leur substitue des conditions reposant sur $\hat{p}$, lesquelles doivent être plus restrictives du fait que $\hat{p}$ est une estimation de p. Une règle simple consiste à vérifier que $n\hat{p}(1-\hat{p}) > 12$. Comme $\hat{p}(1-\hat{p})$ reste inférieur ou égal à $1/4$, cette règle implique $4/n < 1/12$. Alors on peut négliger le terme $\frac{(1{,}96)^2}{n}$ devant 1, donc devant $\hat{p}$ et devant $4\hat{p}(1-\hat{p})$, pour ne conserver finalement que la **formule très classique** :

$$IC_{0,95}(p) \simeq \hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,.$$

Comme pour l'exemple 7.3 on obtient directement cette formule en estimant la variance $np(1-p)$ par $n\hat{P}_n(1-\hat{P}_n)$, car :

$$P\left(-1{,}96 < \frac{n\hat{P}_n - np}{\sqrt{n\hat{P}_n(1-\hat{P}_n)}} < 1{,}96\right) \simeq 0{,}95$$

entraîne :

$$P\left(\hat{P}_n - 1{,}96\sqrt{\frac{\hat{P}_n(1-\hat{P}_n)}{n}} < p < \hat{P}_n + 1{,}96\sqrt{\frac{\hat{P}_n(1-\hat{P}_n)}{n}}\right) \simeq 0{,}95.$$

Cette approche par le théorème central limite coïncide avec l'approche par l'EMV du fait que $I(p) = \frac{1}{p(1-p)}$ pour la loi de Bernoulli. Notons que la *précision absolue* (terme consacré en méthodologie des sondages) : $1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, diminue quand $\hat{p}$ se rapproche de 0 (n étant fixé), mais que la *précision relative :*

$$\frac{1{,}96}{\sqrt{n}}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\hat{p}} = \frac{1{,}96}{\sqrt{n}}\sqrt{\frac{1-\hat{p}}{\hat{p}}}$$

tend vers l'infini.

Exemple 7.5 Soit un sondage auprès d'un échantillon de 1 000 personnes que l'on supposera avoir été sélectionnées au hasard dans la population française des personnes âgées de 18 ans et plus. A la question «Avez-vous une activité sportive au moins une fois par semaine ?» 415 personnes ont répondu affirmativement. Le pourcentage réel dans la population est donc estimé par :

$$0{,}415 \pm 1{,}96\sqrt{\frac{(0,415)(0,585)}{1\,000}},$$

soit 0,415±0,031.

La précision relative est $\frac{0{,}031}{0{,}415}$=0,074 ou 7,4 %. Comme la fonction $\hat{p}(1-\hat{p})$ reste entre 0,24 et 0,25 pour $\hat{p} \in [0{,}40\,;0{,}60]$ on peut retenir que la précision d'un sondage auprès de 1 000 personnes est (au mieux, étant donné les imperfections pratiques) de 0,03 soit, en pourcentage, de 3% pour une proportion située entre 40% et 60%. ■

Un problème classique est celui du **calcul de la taille d'échantillon** pour atteindre une précision absolue donnée et nous prendrons 1% (ou 0,01) pour exemple. Si l'on n'a aucune idée de la valeur de p on peut utiliser le fait que $\hat{p}(1-\hat{p}) \leq \frac{1}{4}$, le maximum étant atteint pour $\hat{p} = \frac{1}{2}$, et la précision sera au pire de $1,96\sqrt{\frac{1}{4n}}$. Donc, en prenant n tel que :

$$1{,}96\sqrt{\frac{1}{4n}} = 0{,}01$$

$$\text{soit } n = \frac{(1{,}96)^2}{4(0{,}01)} \text{ ou } n \simeq 9600,$$

on est sûr d'atteindre la précision souhaitée. Si l'on a une connaissance a priori sur l'ensemble des valeurs plausibles de p (et donc, par assimilation, sur $\hat{p}$) on effectue le même calcul en remplaçant $\hat{p}(1-\hat{p})$ par son maximum sur cet ensemble.

7.4.6 IC sur la différence des paramètres de deux lois de Bernoulli

Soient les deux lois de Bernoulli $\mathcal{B}(p_1)$ et $\mathcal{B}(p_2)$ et **deux échantillons indépendants** issus respectivement de celles-ci, de tailles n_1 et n_2. On s'intéresse à un IC sur $p_1 - p_2$. Les applications sont fréquentes dans les sondages pour comparer les proportions de deux sous-populations dans le choix d'une modalité de réponse à une question donnée. On a donc aussi coutume de parler **d'intervalles de confiance sur la différence de deux proportions** . Pour respecter l'hypothèse d'indépendance des échantillons, les deux sous-populations

doivent être totalement distinctes de façon à donner des sous-échantillons eux-mêmes totalement distincts. Dans ce qui suit, n_1 et n_2 sont supposées fixées, ce qui n'est pas forcément le cas dans cet exemple de sondage où seule la taille globale de l'échantillon n est fixée et les tailles de sous-échantillons sont le résultat du hasard, mais ceci n'a pas vraiment d'incidence sur les résultats établis ci-après (voir à ce propos la note 9.6).

Nous ne donnerons qu'un développement asymptotique qui suppose que les quatre expressions n_1p_1, $n_1(1-p_1)$, n_2p_2 et $n_2(1-p_2)$ soient toutes supérieures ou égales à 5. Les paramètres p_1 et p_2 étant inconnus on peut utiliser en substitution les conditions $n\hat{p}_1(1-\hat{p}_1) > 12$ et $n\hat{p}_2(1-\hat{p}_2) > 12$.

Soit $\hat{P}_1$ et $\hat{P}_2$ les v.a. «proportions de succès» respectives de chaque échantillon. On a alors :

$$\hat{P}_1 \underset{approx}{\rightsquigarrow} \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \quad \text{et} \quad \hat{P}_2 \underset{approx}{\rightsquigarrow} \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{n_2}\right),$$

puis :

$$\hat{P}_1 - \hat{P}_2 \underset{approx}{\rightsquigarrow} \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

Les variances s'additionnent en raison de l'indépendance des deux échantillons et donc des statistiques $\hat{P}_1$ et $\hat{P}_2$. En estimant les variances par $\hat{P}_1(1-\hat{P}_1)/n_1$ et $\hat{P}_2(1-\hat{P}_2)/n_2$, puis en centrant et réduisant, on a :

$$P\left(-1{,}96 \leq \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}} \leq 1{,}96\right) \simeq 0{,}95\ ,$$

ce qui pivote immédiatement pour isoler $p_1 - p_2$ et donne finalement la formule :

$$IC_{0,95}(p_1 - p_2) = (\widehat{p}_1 - \widehat{p}_2) \pm 1{,}96\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

On trouvera dans les exercices des applications illustrant l'intérêt de cette formule.

Signalons, sans développer, qu'il existe une procédure exacte pour les petits échantillons fondée sur la procédure exacte de test correspondante (voir section 9.7.6). Il existe de même une procédure pour le cas d'échantillons appariés.

7.5 IC par la méthode des quantiles

Nous présentons cette méthode, même si elle n'est pas d'un usage très répandu, parce qu'elle est à la base des tables et abaques donnant des IC pour

des petits échantillons, notamment pour la loi de Bernoulli et pour la loi de Poisson. Nous exposerons la démarche dans le cas continu qui est plus simple et indiquerons son adaptation au cas discret.

La méthode exige que l'on dispose d'un estimateur T de θ dont l'expression de la densité $f_T(t;\theta)$ est connue. Il sera bien sûr avantageux que cet estimateur soit de bonne qualité. Il est également nécessaire que la fonction de répartition $F_T(t;\theta)$ soit, pour t fixé et quelconque, une fonction strictement monotone de θ et nous supposerons qu'elle soit, par exemple, strictement décroissante. Cela signifie que le graphe de la densité se déplace vers la droite quand θ augmente.

Définissons la fonction $t_{0,025}(\theta)$ qui à chaque valeur de θ associe le quantile d'ordre 0,025 de la loi correspondante. Cette fonction est strictement croissante. En effet, pour $\theta' > \theta$, on a :

$$F_T(t_{0,025}(\theta);\theta) = 0,025 > F_T(t_{0,025}(\theta);\theta')$$

et $t_{0,025}(\theta)$ est donc un quantile d'ordre inférieur à 0,025 pour la loi correspondant à θ'. Ainsi $t_{0,025}(\theta')$, le quantile d'ordre 0,025 de cette dernière, est supérieur à $t_{0,025}(\theta)$. Définissons de même la fonction $t_{0,975}(\theta)$ qui à chaque valeur de θ associe le quantile d'ordre 0,975 de la loi correspondante. Cette fonction est également strictement croissante.

Ayant observé $T = t$, l'IC à 95% pour θ par la méthode des quantiles est $[\theta_1, \theta_2]$ où θ_1 est tel que $t_{0,975}(\theta_1) = t$ et θ_2 est tel que $t_{0,025}(\theta_2) = t$ (voir la figure 7.1). En d'autres termes, θ_1 est la valeur dans Θ dont la loi correspondante a pour quantile d'ordre 0,975 la valeur observée t et de même pour θ_2 avec le quantile d'ordre 0,025. Les fonctions $t_{0,975}(\theta)$ et $t_{0,025}(\theta)$ étant monotones on peut écrire $\theta_1 = t^{-1}_{0,975}(t)$ et $\theta_2 = t^{-1}_{0,025}(t)$.

Montrons qu'on a bien en amont une procédure d'intervalle de confiance de niveau 0,95. Considérons donc l'intervalle aléatoire $[t^{-1}_{0,975}(T)\ ,\ t^{-1}_{0,025}(T)]$. On a :

$$P_\theta\left(t^{-1}_{0,975}(T) < \theta < t^{-1}_{0,025}(T)\right) = P_\theta\left(t_{0,025}(\theta) < T < t_{0,975}(\theta)\right)$$

ce qui, quel que soit θ, est, par définition des quantiles, égal à 0,95. L'application de cette procédure à la loi $\mathcal{U}[0,\theta]$ est proposée en exercice.

Cas d'une loi discrète

Intéressons-nous maintenant à une famille de lois discrètes pour laquelle la statistique T sera également discrète de fonction de probabilité $p_T(x;\theta)$ et de fonction de répartition $F_T(t;\theta)$ strictement décroissante en θ comme précédemment. De plus, pour simplifier les écritures nous supposerons que pour tout θ l'ensemble des valeurs possibles de T est $\mathbb{N}$. La procédure ci-dessus n'est plus possible car, en raison des sauts de discontinuité de $F_T(t;\theta)$, on ne peut pas systématiquement associer à un θ donné un quantile d'ordre exactement

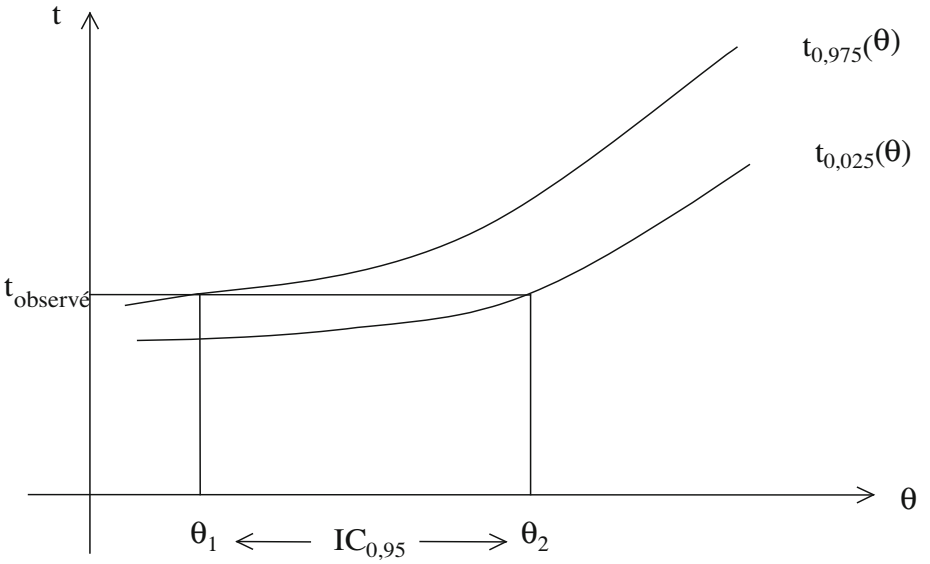


Figure 7.1 - Intervalle de confiance par la méthode des quantiles.

égal à 0,025 ou 0,975. En revanche, $p_T(t;\theta)$ étant généralement continue en θ, pour t donné on peut toujours trouver θ_1 tel que :

$$F_T(t;\theta_1) = \sum_{x=0}^{t} p_T(x;\theta_1) = 0,975$$

et de même on peut trouver θ_2 tel que $F_T(t;\theta_2) = 0,025$.

En d'autres termes, pour tout $t \in \mathbb{N}$, $t_{0,975}^{-1}(t)$ et $t_{0,025}^{-1}(t)$ sont définis. En fait on montre (voir la note 7.6) que si l'on veut garantir pour tout θ une probabilité **au moins égale à 0,95** (on dit alors que la procédure est conservatrice) on doit prendre, ayant observé t, comme intervalle de confiance :

$$IC_{0,95}(\theta) = [\theta_1, \theta_2]$$

où :

$$\theta_1 \text{ est tel que } \sum_{x=0}^{t-1} p_T(x;\theta_1) = 0,975\,,$$

$$\theta_2 \text{ est tel que } \sum_{x=0}^{t} p_T(x;\theta_2) = 0,025.$$

Exemple 7.6 Soit la famille des lois de Poisson et un échantillon de taille 7 issu de la loi $\mathcal{P}(\lambda)$ où λ est inconnu. La statistique efficace (voir section 6.6.3) $T = \sum_{i=1}^{7} X_i$ suit une loi $\mathcal{P}(7\lambda)$. Le graphe de la fonction de probabilité d'une loi de Poisson $\mathcal{P}(\rho)$ se déplaçant vers la droite quand ρ augmente, la fonction de répartition $F_T(t;\rho)$ doit être strictement décroissante en ρ pour t fixé. Montrons-le rigoureusement. On a :

$$F_T(t;\rho) = \sum_{x=0}^{t} \frac{e^{-\rho}\rho^x}{x!}$$

$$\begin{aligned}\frac{\partial}{\partial\rho}F_T(t;\rho) &= -e^{-\rho} + \sum_{x=1}^{t} \frac{xe^{-\rho}\rho^{x-1} - e^{-\rho}\rho^x}{x!} \\ &= \sum_{x=1}^{t} \frac{e^{-\rho}\rho^{x-1}}{(x-1)!} - \sum_{x=0}^{t} \frac{e^{-\rho}\rho^x}{x!} \\ &= -\frac{e^{-t}\rho^t}{t!} \quad < 0.\end{aligned}$$

Supposons que l'on ait observé un total des observations $\sum_{i=1}^{7} x_i = 18$. L'IC pour λ à 95% est donné par $[\lambda_1, \lambda_2]$ où :

$$\lambda_1 \text{ est tel que } \sum_{x=0}^{17} \frac{e^{-7\lambda_1}(7\lambda_1)^x}{x!} = 0{,}975\ ,$$
$$\lambda_2 \text{ est tel que } \sum_{x=0}^{18} \frac{e^{-7\lambda_2}(7\lambda_2)^x}{x!} = 0{,}025.$$

En recourant à un logiciel mathématique on trouve les solutions $7\lambda_1$ =10,7 et $7\lambda_2$ =28,4 soit finalement :

$$IC_{0,95}(\lambda) = \left[\frac{10{,}7}{7}; \frac{28{,}4}{7}\right] = [1{,}53\ ;\ 4{,}06]\,.$$

En exercice 7.4, on montrera comment on peut également résoudre les deux équations ci-dessus à l'aide d'une table des lois du Khi-deux.

Comparons ce résultat avec celui de la formule asymptotique de l'exemple 7.3 :

$$\begin{aligned}IC_{0,95}(\lambda) &= \frac{18}{7} \pm 1{,}96\sqrt{\frac{18/7}{7}} \\ &= [1{,}38\ ;\ 3{,}76].\end{aligned}$$

Ce dernier est un peu plus étroit mais, étant approché, on ne peut garantir le niveau 0,95 , à savoir que la probabilité de couverture de λ par la méthode asymptotique n'est pas nécessairement égale ou supérieure à 0,95 pour tout

λ. Si l'on avait trouvé $\sum_{i=1}^{7} x_i = 50$, la procédure conservatrice aurait donné l'intervalle [5,29 ; 9,41] et la procédure asymptotique [5,16 ; 9,12]. De fait, les procédures se rapprochent quand λ augmente. ■

Dans les tables ou les logiciels statistiques élaborés on obtient directement les valeurs $n\lambda_1$ et $n\lambda_2$ pour les valeurs de $\sum_{i=1}^{n} x_i$ de 0 à 50. Au-delà, on peut utiliser l'IC asymptotique.

L'approche est identique pour un IC sur le paramètre p de la **loi de Bernoulli** pour laquelle statistique T est également la somme $\sum_{i=1}^{n} X_i$ (le nombre total de succès), de loi $\mathcal{B}(n, p)$. Ici il faut tenir compte à la fois de la valeur de $\sum_{i=1}^{n} x_i$ et de celle de n. C'est pourquoi les bornes de l'IC sont données sous forme d'abaques. Un exemple pour cette loi est proposé dans les exercices.

Note 7.6 Montrons que la procédure adoptée pour le cas discret est conservatrice. Pour la valeur t observée, la borne θ_1 est telle que $F_T(t-1;\theta_1) = 0,975$. Calculons pour un θ quelconque $P_\theta(\theta_1(T) \leq \theta)$. L'événement $\{\theta_1(T) \leq \theta\}$ est la partie A de $\mathbb{N}$ définie par $A = \{t \,|\, \theta_1(t) \leq \theta\}$ ou, puisque $F_T(u;\theta)$ est strictement décroissante en θ pour u fixé quelconque, de façon équivalente,

$$A = \{t \mid F_T(t-1;\theta_1) \geq F_T(t-1;\theta)\} = \{t \mid F_T(t-1;\theta) \leq 0,975\}.$$

A est donc constitué de toutes les valeurs de t de 0 à t_0 où t_0 est la première valeur telle que $F_T(t_0;\theta) > 0,975$. D'où $P(A) = F_T(t_0;\theta) > 0,975$.

Par la même argumentation on peut montrer que $P_\theta(\theta \leq \theta_2(T)) > 0,975$ pour tout θ, d'où $P_\theta(\theta_1(T) < \theta < \theta_2(T)) > 0,975 - 0,025 = 0,95$.

7.6 Approche bayésienne

Dans cette approche nous avons vu en section 6.8 que le paramètre θ avait un statut de variable aléatoire. A la notion d'intervalle de confiance de niveau γ on substitue la notion d'intervalle de probabilité γ sur la loi *a posteriori* de θ. On encadrera donc simplement θ avec les quantiles d'ordre 0,025 et 0,975 sur cette loi.

Exemple 7.7 Reprenons l'exemple 6.23 de l'estimation du paramètre p d'une loi de Bernoulli avec une loi *a priori* $\mathcal{U}[0,1]$. La loi *a posteriori* est une loi $Beta(s, n-s)$, où s est le nombre de succès, dont on peut trouver les quantiles dans les logiciels statistiques ou dans les tables. Supposons que pour $n = 20$ répétitions on ait observé $s = 8$ succès. On lit dans une table les quantiles d'ordres 0,025 et 0,975 de la loi $Beta(8;12)$: 0,22 et 0,62 respectivement. D'où l'intervalle de probabilité 0,95 pour p : [0,22 ; 0,62], à comparer avec l'intervalle [0,19 ; 0,64] indiqué par la méthode des quantiles dans l'exercice 7.5. ■

7.7 Notions d'optimalité des IC

Le premier critère d'optimalité est celui de largeur minimale des intervalles produits par la procédure. C'est d'ailleurs cette idée qui nous a conduit à choisir les quantiles de façon symétrique sur les extrémités de la distribution (soit les quantiles d'ordre $\frac{1-\gamma}{2}$ et $\frac{1+\gamma}{2}$ pour un IC de niveau $1-\gamma$) dans la mesure où l'on obtient ainsi l'intervalle le plus court lorsque la distribution concernée est symétrique avec un seul mode (voir note 7.1). Nous définissons ci-après la notion de procédure de largeur minimale qui explicite le fait que l'IC doit être le plus court quelle que soit la réalisation $(x_1, x_2, \cdots, x_n)$.

Définition 7.4 *Une procédure d'IC est dite de* ***largeur minimale*** *au niveau γ si la largeur de son IC de niveau γ : $[t_1(x_1, x_2, \cdots, x_n), t_2(x_1, x_2, \cdots, x_n)]$, est inférieure à celle de tout autre IC dérivé d'une procédure de niveau égal ou supérieur à γ, et ceci pour toute réalisation $(x_1, x_2, \cdots, x_n)$.*

Il n'est évidemment pas aisé de dégager une telle procédure. Cela ne peut être fait que dans quelques cas simples mais, en général, une telle procédure n'existera pas. Un critère plus faible consiste à raisonner non pas pour toute réalisation, mais par rapport à l'espérance mathématique de la largeur de l'intervalle $E_\theta[t_2(X_1, X_2, \cdots, X_n) - t_1(X_1, X_2, \cdots, X_n)]$ quel que soit θ.

Cependant il existe un résultat asymptotique intéressant concernant la procédure de la section 7.3 reposant sur l'estimateur du maximum de vraisemblance, résultat qui découle de l'optimalité asymptotique de celui-ci (voir proposition 6.11) : sous certaines conditions de régularité cette procédure fournira une largeur d'intervalle qui, en espérance mathématique, tendra à être minimale pour $n \to \infty$.

Un autre critère d'optimalité est fourni par la notion de procédure uniformément plus précise. Cette notion semblera ici quelque peu complexe mais elle deviendra plus claire lorsqu'aura été vue la notion duale de test uniformément plus puissant au chapitre 9.

Définition 7.5 *Une procédure d'IC (T_1^*, T_2^*) est dite* ***uniformément plus précise*** *au niveau γ qu'une procédure (T_1, T_2) si, étant toutes deux de niveau γ, on a :*

$$P_\theta(T_1^* < \theta' < T_2^*) \leq P_\theta(T_1 < \theta' < T_2)$$

pour tout $\theta \in \Theta$ et pour tout $\theta' \in \Theta$ différent de θ, l'inégalité étant stricte pour au moins une valeur de θ.

En d'autres termes, la procédure sera plus précise si la probabilité d'encadrer une valeur θ' autre que la vraie valeur de θ reste plus faible. L'objectif sera alors de rechercher, si elle existe, la procédure uniformément la plus précise (en anglais : *uniformly most accurate* ou UMA) parmi l'ensemble des procédures de niveau égal (ou supérieur) à γ.

Enfin, on peut souhaiter d'une procédure que la largeur des IC fournis tende vers 0 quand la taille de l'échantillon s'accroît.

Définition 7.6 *Soit une procédure d'IC fondée pour $n \in \mathbb{N}^*$ sur l'intervalle aléatoire $[T_{1,n}, T_{2,n}]$. On dit que cette procédure est* ***convergente*** *en probabilité si la suite $\{T_{2,n} - T_{1,n}\}$ est telle que :*

$$T_{2,n} - T_{1,n} \quad \xrightarrow[n\to\infty]{p} \quad 0\,.$$

Étant donné que, pour chaque n, l'intervalle $[T_{1,n}, T_{2,n}]$ doit contenir la vraie valeur de θ avec une forte probabilité, cet intervalle se réduira, à la limite, à cette valeur. Prenons par exemple l'intervalle :

$$\left[\bar{X}_n - t_{0,975}^{(n-1)}\,\frac{S_n}{\sqrt{n}}\,,\ \bar{X}_n + t_{0,975}^{(n-1)}\,\frac{S_n}{\sqrt{n}}\right]$$

de la procédure classique pour la moyenne μ de la loi $\mathcal{N}(\mu\,,\,\sigma^2)$ vue en section 7.4.1. La largeur de l'intervalle est $2\,t_{0,975}^{(n-1)}\,\frac{S_n}{\sqrt{n}}$. Comme S_n^2 converge en probabilité vers σ^2, S_n converge vers σ et la largeur converge en probabilité vers 0. Par ailleurs $\bar{X}_n$ converge vers μ et cet intervalle se réduit à μ à l'infini. Cette propriété de convergence est vérifiée pour tous les intervalles classiques que nous avons présentés. Elle est également vraie pour la procédure asymptotique par l'EMV (voir section 7.3) dans la mesure où $I(\hat{\theta}_n^{MV})$ converge vers $I(\theta_0)$ où θ_0 est la vraie valeur de θ.

7.8 Région de confiance pour un paramètre de dimension $k > 1$

Pour simplifier nous prendrons $k = 2$, l'extension à k quelconque ne présentant pas de difficultés particulières. Soit donc $\theta = (\theta_1, \theta_2)$ le paramètre inconnu. Le problème est maintenant de déterminer une région aléatoire du plan qui contienne θ avec une probabilité donnée quel que soit θ.

Supposons d'abord que l'on sache construire séparément pour chaque composante une procédure d'IC de niveau γ et soit I_1 et I_2 les intervalles aléatoires correspondants. Pour tout $\theta \in \Theta \subseteq \mathbb{R}^2$, on a alors :

$$P_\theta(\theta_j \in I_j) = \gamma, \quad j = 1, 2,$$

en admettant, également pour simplifier, que la probabilité γ est exactement atteinte pour chaque θ. Considérons la région aléatoire constituée du rectangle $I_1 \times I_2$. Pour θ fixé on a :

$$P_\theta(\theta \in I_1 \times I_2) = P_\theta(\theta_1 \in I_1, \theta_2 \in I_2) = P_\theta((\theta_1 \in I_1) \cap (\theta_2 \in I_2))$$

qui sera toujours inférieur à γ (sauf cas très particulier où l'un des événements implique l'autre). Généralement ces deux événements seront dépendants (du fait qu'ils reposent sur les mêmes observations) et il sera difficile de déterminer cette probabilité et donc, en prenant la valeur minimale quand θ décrit Θ, de connaître le niveau de confiance exact associé à la procédure consistant à prendre le rectangle au croisement de deux intervalles. Toutefois montrons que l'on peut donner une borne inférieure pour cette probabilité. Pour ce faire, posons $\alpha = 1-\gamma$ qui correspond au risque d'erreur de la procédure pour chaque composante.

Soit E_1 et E_2 deux événements quelconques. Le complémentaire de $E_1 \cap E_2$ est $\overline{E}_1 \cup \overline{E}_2$. Par ailleurs (voir section 1.1) :

$$P(\overline{E}_1 \cup \overline{E}_2) = P(\overline{E}_1) + P(\overline{E}_2) - P(\overline{E}_1 \cap \overline{E}_2) \leq P(\overline{E}_1) + P\overline{E}_2).$$

D'où l'inégalité générale :

$$P(E_1 \cap E_2) \geq 1 - \left[P\overline{E}_1) + P(\overline{E}_2)\right].$$

Appliquant celle-ci aux événements $(\theta_1 \in I_1)$ et $(\theta_2 \in I_2)$ on en déduit :

$$P_\theta((\theta_1 \in I_1) \cap (\theta_2 \in I_2)) \geq 1 - 2\alpha.$$

Ainsi, si l'on vise un niveau de confiance $1-\alpha$, on peut le garantir en prenant pour chaque composante un niveau de confiance $1-\frac{\alpha}{2}$. (Pour le niveau courant de 0,95 on utilisera les IC de niveau 0,975 sur chaque composante). Pour une dimension k quelconque la méthode ci-dessus s'applique en prenant un niveau $1-\frac{\alpha}{k}$ sur chaque composante.

Toutefois cette procédure peut s'avérer très conservatrice, au sens où le niveau réel sera supérieur et la région du plan sera donc plus vaste que nécessaire. Il n'est donc pas inutile de rechercher de façon directe une région de niveau γ. Nous illustrons une approche exacte pour le paramètre (μ, σ^2) de la loi $\mathcal{N}(\mu\,, \sigma^2)$.

Nous avons vu (proposition 5.3) que $\bar{X}$ et S^2 sont indépendants, ce qui implique que $\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}}$ et $\frac{(n-1)S^2}{\sigma^2}$ le sont également. Ces deux dernières v.a. étant, respectivement, de lois $\mathcal{N}(0\,;1)$ et $\chi^2(n-1)$ on a, quel que soit (μ, σ^2) :

$$P\left(-2{,}24 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 2{,}24\right) = 0{,}975$$

$$P\left(\chi^{2\,(n-1)}_{0{,}0125} < \frac{(n-1)S^2}{\sigma^2} < \chi^{2\,(n-1)}_{0{,}9875}\right) = 0{,}975$$

et la probabilité que ces deux événements aient lieu simultanément est donc $(0,975)^2 \simeq 0,95$. Ainsi une région de confiance de niveau 0,95 est obtenue en

prenant l'ensemble des points (μ, σ^2) du plan tels que :

$$\begin{cases} -2{,}24 < \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 2{,}24 \\ \chi^{2\,(n-1)}_{0{,}0125} < \dfrac{(n-1)s^2}{\sigma^2} < \chi^{2\,(n-1)}_{0{,}9875} \end{cases}$$

ou, de façon équivalente :

$$\begin{cases} \sigma^2 - \dfrac{n}{(2{,}24)^2}(\mu - \bar{x})^2 > 0 \\ \dfrac{(n-1)s^2}{\chi^{2\,(n-1)}_{0{,}9875}} < \sigma^2 < \dfrac{(n-1)s^2}{\chi^{2\,(n-1)}_{0{,}0125}} \end{cases} .$$

La première inégalité correspond, en coordonnées (x, y), à l'intérieur de la parabole $y = a(x - \bar{x})^2$ centrée sur $\bar{x}$ où a vaut $n/(2{,}24)^2$. La seconde découpe une tranche de cet intérieur entre les droites horizontales d'équations $y = \frac{(n-1)s^2}{\chi^{2\,(n-1)}_{0{,}9875}}$ et $y = \frac{(n-1)s^2}{\chi^{2\,(n-1)}_{0{,}0125}}$ comme indiqué sur la figure 7.2 .

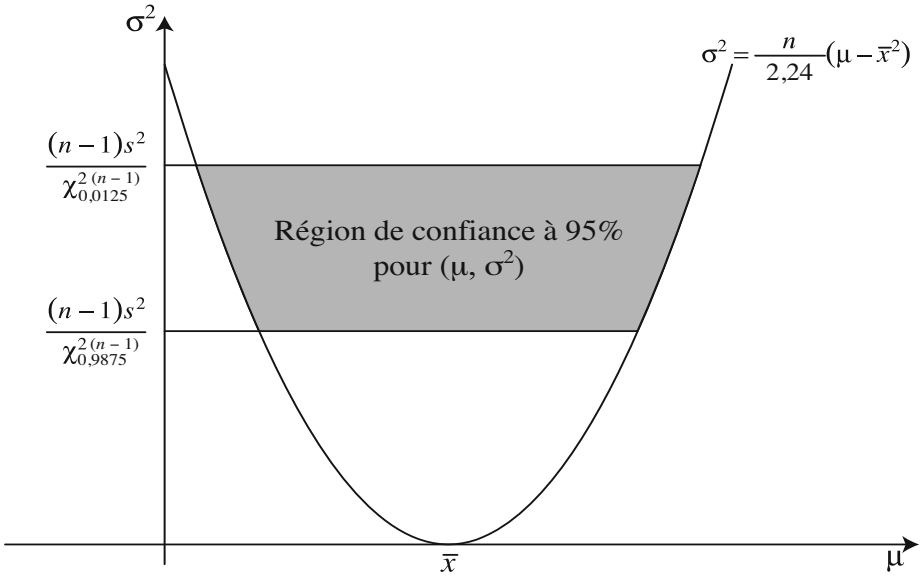


Figure 7.2 - Région de confiance pour le paramètre (μ, σ^2) d'une loi de Gauss.

Signalons brièvement les éléments permettant d'obtenir des régions de confiance approximatives dans $\mathbb{R}^k$ à partir des propriétés du vecteur estimateur du MV. Nous avons indiqué en fin de section 6.7.4 que $\sqrt{n}(\hat{\theta}^{MV} - \theta)$ converge en

loi vers une loi normale à k dimensions de vecteur des moyennes nul et de matrice des variances-covariances $[\mathbb{I}(\theta))]^{-1}$. Comme la matrice $\mathbb{I}(\theta)$ est symétrique et définie strictement positive il existe une matrice symétrique et définie strictement positive dont le carré est égal à $\mathbb{I}(\theta)$ et nous la désignons par $\mathbb{I}(\theta)^{\frac{1}{2}}$. Ceci est également applicable à $[\mathbb{I}(\theta)]^{-1}$ et $\mathbb{I}(\theta)^{-\frac{1}{2}}$ est l'inverse de $\mathbb{I}(\theta)^{\frac{1}{2}}$, i.e. $\mathbb{I}(\theta)^{\frac{1}{2}}\mathbb{I}(\theta)^{-\frac{1}{2}} = \mathbf{I}_k$, la matrice identité d'ordre k.

Posons $X = \sqrt{n}(\hat{\theta}^{MV} - \theta)$ et soit $Y = \mathbb{I}(\theta)^{\frac{1}{2}}X$. Selon la proposition 3.13, $\mathbb{E}(Y) = \mathbb{I}(\theta)^{\frac{1}{2}}\mathbb{E}(X) = 0$ et :

$$\mathbb{V}(Y) = \mathbb{I}(\theta)^{\frac{1}{2}}\mathbb{V}(X)\,\mathbb{I}(\theta)^{\frac{1}{2}} = \mathbb{I}(\theta)^{\frac{1}{2}}\mathbb{I}(\theta)^{-1}\mathbb{I}(\theta)^{\frac{1}{2}} = \mathbf{I}_k.$$

Ainsi $\sqrt{n}\,\mathbb{I}(\theta)^{\frac{1}{2}}(\hat{\theta}^{MV} - \theta)$ a une loi asymptotique $\mathcal{N}(0\,, \mathbf{I}_k)$, c'est-à-dire que toutes les composantes de ce vecteur aléatoire tendent à être indépendantes et de loi $\mathcal{N}(0\,;1)$ (voir les développements analogues en section 3.9). Par conséquent la somme des carrés de ses composantes, égale à :

$$\left[\sqrt{n}\,\mathbb{I}(\theta)^{\frac{1}{2}}(\hat{\theta}^{MV} - \theta)\right]^t \left[\sqrt{n}\,\mathbb{I}(\theta)^{\frac{1}{2}}(\hat{\theta}^{MV} - \theta)\right] = n(\hat{\theta}^{MV} - \theta)^t\,\mathbb{I}(\theta)\,(\hat{\theta}^{MV} - \theta)\,,$$

suit approximativement une loi du khi-deux à k degrés de liberté. En remplaçant, en deuxième approximation, $\mathbb{I}(\theta)$ par $\mathbb{I}(\hat{\theta}^{MV})$ et en passant à la réalisation de $\hat{\theta}^{MV}$, l'inéquation en θ :

$$n(\hat{\theta}^{MV} - \theta)^t\,\mathbb{I}(\hat{\theta}^{MV})\,(\hat{\theta}^{MV} - \theta) \leq \chi^2_{0,95}(k)$$

définit l'intérieur d'un ellipsoïde centré sur $\hat{\theta}^{MV}$ qui est une région de confiance de niveau approximatif 0,95.

Appliquant ceci à (μ, σ^2) dans le cas gaussien on a (voir exemple 6.18) :

$$\mathbb{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \quad \text{et} \quad [\mathbb{I}(\mu, \sigma^2)]^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

D'où la région de confiance au niveau 0,95 (où ici la substitution $\mathbb{I}(\bar{x}, \tilde{s}^2)$ pour $\mathbb{I}(\mu, \sigma^2)$ n'est pas nécessaire) :

$$\frac{n}{\sigma^2}(\mu - \bar{x})^2 + \frac{n}{2\sigma^4}(\sigma^2 - \tilde{s}^2)^2 \leq \chi^2_{0,95}(2) = 5{,}99$$

qui correspond à l'intérieur d'une ellipse centrée sur l'estimation du MV : $(\bar{x}, \tilde{s}^2)$. En fait, on montre que cette région est plus intéressante que celle obtenue plus haut car elle est (en espérance mathématique, et pour n pas trop petit) de surface inférieure.

Grâce à la résolution numérique vue en section 7.3 permettant d'accéder à $I(\hat{\theta}^{MV})$, les logiciels peuvent, pour $k = 2$, tracer les ellipses contenant le paramètre à un niveau de confiance donné.

7.9 Intervalles de confiance et tests

Nous verrons en section 9.8 qu'il y a une dualité entre une procédure d'IC et une procédure de test. Comme il est généralement plus facile d'élaborer une procédure de test, nous montrerons comment, à partir de celle-ci, construire un intervalle de confiance. Cela permettra de couvrir des situations encore plus diverses que celles envisageables par les approches directes du présent chapitre.

Pour approfondir la théorie des intervalles de confiance (et, plus généralement, la théorie de l'estimation et des tests) on pourra consulter le livre de Cox et Hinkley (1979) ou celui de Shao (1999).

7.10 Exercices

Exercice 7.1 Pour la loi de Gauss $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu comparer la largeur de l'IC obtenu pour μ à celle de l'IC obtenu pour σ quand n est grand.

Aide : pour σ on partira de la formule de la note 7.5 et on utilisera l'approximation de la loi du χ^2 par une loi de Gauss.

Exercice 7.2 Soit la loi de Gauss $\mathcal{N}(\mu, \sigma^2)$ où μ est connu – on prendra $\mu = 0$ – et σ^2 inconnu. Donner un IC pour σ^2.

Aide : montrer que $\sum_{i=1}^{n} \frac{X_i^2}{\sigma^2}$ est une fonction pivot.

Exercice 7.3 (méthode des quantiles) Donner un IC à 95% pour θ de la loi $\mathcal{U}[0, \theta]$ en utilisant la loi de la statistique exhaustive minimale $X_{(n)}$.

Exercice 7.4 (méthode des quantiles) * Soit $F_{\chi^2}(x; 2n)$ la fonction de répartition de la loi du khi-deux à $2n$ degrés de liberté. Montrer que :

$$F_{\chi^2}(x; 2n) = 1 - \sum_{k=0}^{n-1} \frac{e^{-\frac{x}{2}} (\frac{x}{2})^k}{k!}.$$

Aide : on calculera l'intégrale de la densité en intégrant par parties et en exploitant la relation de récurrence obtenue.

Soit $F_P(x; \lambda)$ la fonction de répartition de la loi de Poisson $\mathcal{P}(\lambda)$. Montrer que $F_P(x; \lambda) = 1 - F_{\chi^2}(2\lambda; 2x + 2)$ et que, grâce à cette relation, on peut résoudre immédiatement les deux équations de la méthode des quantiles pour la loi de Poisson, à l'aide d'une table du khi-deux. Consulter une telle table pour vérifier le résultat de l'exemple 7.6.

Exercice 7.5 (méthode des quantiles) Soit un échantillon de la loi de Bernoulli de taille 20 pour lequel on a observé $\sum_{i=1}^{n} X_i = 8$ (soit 8 succès au cours de 20 répétitions). Montrer, en résolvant les équations de la méthode

des quantiles au moyen d'un logiciel mathématique, qu'on obtient un IC à 95% égal à [0,19 ; 0,64]. Comparer à l'IC de la formule asymptotique.

Exercice 7.6 Soit la loi mère $\mathcal{N}(\mu\,,\sigma^2)$ où μ est inconnu mais σ^2 est connu. Dans une approche bayésienne on se donne une loi *a priori* $\mathcal{N}(\mu_0\,,\sigma_0^2)$ pour μ. Montrer que la loi *a posteriori* de μ est gaussienne de moyenne :

$$\frac{\sigma_0^2\,\bar{x} + \sigma^2\mu_0/n}{\sigma_0^2 + \sigma^2/n}$$

et de variance :

$$\frac{\sigma_0^2\sigma^2/n}{\sigma_0^2 + \sigma^2/n}$$

où $\bar{x}$ est la moyenne observée sur un échantillon de taille n. En déduire un intervalle de probabilité de niveau 0,95 et le comparer à l'IC classique avec σ^2 connu (voir fin de section 7.4.1). Montrer que les deux intervalles sont équivalents quand $n \to \infty$.

Exercice 7.7 Montrer que la procédure d'IC proposée en section 7.4.5 pour le paramètre p d'une loi de Bernoulli est convergente en probabilité.

Exercices appliqués[1]

Exercice 7.8 On veut estimer le rendement d'un engrais pour la culture du blé. Sur douze parcelles expérimentales, on a trouvé les rendements suivants en tonnes par hectare :

7.7 8.4 7.8 8.2 7.9 8.5 8.4 8.2 7.6 7.8 8.4 8.3

Donner un intervalle de confiance à 95% pour le rendement moyen de l'engrais (on supposera que le rendement à l'hectare est une v.a. gaussienne).

Exercice 7.9 Un hôpital souhaite estimer le coût moyen d'un patient, sachant que le coût par jour est de 200 euros. Pour un échantillon aléatoire de 500 patients on a observé une durée de séjour moyenne de 5,4 jours avec un écart-type de 3,1 jours. Donner un intervalle de confiance à 90 % pour la durée moyenne de séjour d'un patient et en déduire un intervalle pour le coût moyen d'un patient.

Exercice 7.10 Une société d'assurance doit évaluer, en fin d'année, la provision à faire au bilan pour les sinistres en cours n'ayant pas encore fait l'objet d'un règlement. Elle sélectionne au hasard 200 dossiers qui sont évalués en moyenne à 9 944 euros, l'écart-type des valeurs étant égal à 1 901 euros. Sachant que 11 210 dossiers sont en cours, donner un intervalle de confiance sur la provision totale à effectuer.

[1] Un ou deux de ces exercices appliqués sont des emprunts dont nous avons perdu la source. Nous nous en excusons auprès des involontaires contributeurs.

Exercice 7.11 Pour évaluer le nombre de mots d'un livre on tire 20 pages au hasard et on y compte le nombre de mots. On trouve, pour les 20 valeurs, une moyenne de 614 mots et un écart-type de 26 mots. Donner un intervalle de confiance à 95 % pour le nombre total de mots du livre sachant qu'il a 158 pages (on admettra que l'approximation gaussienne est satisfaisante).

Exercice 7.12 On souhaite évaluer le gain de consommation obtenu avec un nouveau carburant pour automobile. Un test en laboratoire est effectué sur 20 moteurs du même type. Dix moteurs sont alimentés en carburant traditionnel et donnent sur une durée donnée une consommation moyenne de 10,8 litres avec un écart-type de 0,21 litre. Pour les dix autres moteurs le nouveau carburant est utilisé et l'on observe une consommation moyenne de 10,3 litres avec un écart-type de 0,18 litre. Donner un intervalle de confiance à 90 % sur le gain moyen procuré par le nouveau carburant (on supposera que les approximations gaussiennes sont satisfaisantes).

Exercice 7.13 Un stock comporte 10 000 pièces. Pour évaluer le nombre de pièces défectueuses dans le stock on tire au hasard 400 pièces dont on constate que 45 sont défectueuses.

Donner un intervalle de confiance à 99 % pour le nombre total de pièces défectueuses.

Exercice 7.14 Un sondage auprès de 1 500 ménages tirés au hasard dans la population française a indiqué que 20 % de ceux-ci prévoient d'acheter une nouvelle voiture dans les douze prochains mois. Estimer par un intervalle de confiance à 95 % le pourcentage de ménages de la population française prévoyant d'acheter une nouvelle voiture dans les douze mois.

Exercice 7.15 On veut évaluer la différence des proportions de pièces défectueuses dans deux procédés de fabrication différents. Pour cela on tire au hasard 1 000 pièces réalisées selon le premier procédé. Les ayant testées on en a trouvé 86 défectueuses.

On opère de même pour 800 pièces réalisées selon le deuxième procédé et on en trouve 92 défectueuses.

Donner un intervalle de confiance sur la différence des proportions de pièces défectueuses dans les deux procédés.

Exercice 7.16 Dans une ville on donne la répartition du nombre de jours sans accident, avec un accident, etc. parmi 50 jours d'observation au cours d'une même année :

Nbre accidents	0	1	2	3	4
Nbre jours	21	18	7	3	1

On suppose que le nombre d'accidents par jour suit une loi de Poisson.

Donner un intervalle de confiance de niveau 0,95 pour le nombre moyen d'accidents par jour (on utilisera une approximation asymptotique).

Exercice 7.17 Dix bouteilles d'eau minérale provenant d'une source donnée sont analysées. On relève les taux de nitrates suivants, en mg/l :

3,61 3,56 3,67 3,56 3,64 3,62 3,44 3,52 3,55 3,52

Donner un intervalle de confiance à 95% pour l'écart-type du taux de nitrates dans les bouteilles produites (on supposera ce taux gaussien).

Chapitre 8

Estimation non paramétrique et estimation fonctionnelle

8.1 Introduction

Nous considérons maintenant que la loi mère ne fait pas partie d'une famille paramétrable de lois, c'est-à-dire que nos connaissances sur la nature de cette loi sont beaucoup plus floues, ce qui correspond d'ailleurs souvent plus à la réalité, notamment lorsqu'il s'agit d'un sondage dans une population. Tout au plus ferons-nous ici ou là l'hypothèse que sa fonction de répartition, ou sa densité (cas continu), ou sa fonction de probabilité (cas discret) répond à des conditions de régularité, principalement la dérivabilité et l'existence de moments jusqu'à un certain ordre.

Il ne peut donc plus s'agir ici d'estimer un paramètre qui déterminerait totalement la loi et par suite toute caractéristique de celle-ci. Dès lors deux orientations sont possibles. Soit on s'intéresse uniquement à quelques valeurs caractéristiques de la loi (ou de la population dans la situation de sondage) : moyenne, variance ou écart-type, médiane ou tout autre quantile, et dans ce cas nous sommes dans un contexte d'*estimation non paramétrique ponctuelle*. Soit, ce qui est nouveau par rapport à l'estimation paramétrique, on veut estimer la loi dans sa globalité par sa fonction de répartition ou sa densité, ou sa fonction de probabilité (quoique, on le verra, le cas discret soit peu concerné) et l'on parle alors d'*estimation fonctionnelle*. Pour illustrer cette deuxième orientation, disons déjà que l'histogramme utilisé en statistique descriptive est une façon rudimentaire d'approcher la densité de la loi, dont nous étudierons d'ailleurs l'efficacité.

Nous commençons par l'estimation ponctuelle en nous bornant aux caractéristiques mentionnées ci-dessus. Nous reprendrons certains résultats des chapitres précédents dont nous avons pu dire qu'ils étaient en fait de portée générale et non pas limitée au cadre paramétrique. Dans la mesure du possible nous traiterons simultanément le problème de la construction d'un intervalle de confiance.

Dans ce chapitre, comme précédemment, la loi mère sera symbolisée par la v.a. générique X, sa fonction de répartition par F, sa fonction de densité ou de probabilité par f.

8.2 Estimation de la moyenne et de la variance de la loi

8.2.1 Estimation de la moyenne μ

Nous avons vu en section 6.5.1 que les moments empiriques simples (s'ils existent) sont des estimateurs sans biais des moments correspondants de la loi mère et ceci quelle que soit la nature de cette loi. En section 6.5.3 on a vu encore qu'en conséquence de la loi des grands nombres, ces estimateurs sont convergents presque sûrement. Pour que le moment d'ordre k converge il suffit que $E(|X^k|)$ existe (voir proposition 6.1). Cela s'applique évidemment à la moyenne empirique $\overline{X} = \sum_{i=1}^{n} X_i$. Notons que, pour un sondage dans une population finie, ces conditions d'existence des moments sont nécessairement réunies, ceux-ci étant des caractéristiques descriptives de la population dans son ensemble. Pour $\overline{X}$ nous avions directement établi en section 5.2 (proposition 5.1) que (si σ^2 existe) :

$$E(\overline{X}) = \mu \quad \text{et} \quad V(\overline{X}) = \frac{\sigma^2}{n},$$

la première relation reflétant un biais nul et la deuxième montrant directement la convergence en moyenne quadratique.

En conclusion nous utiliserons naturellement la moyenne empirique pour estimer la moyenne de la loi, nous satisfaisant de ces propriétés. Il n'est pas possible de dire si tel est le meilleur choix possible, sauf à imposer des conditions restrictives sur la nature de la loi mère, ce qui n'est pas dans l'esprit de l'estimation non paramétrique.

En section 7.4.1 nous avons noté que, dès lors que n est assez grand, on a :

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \underset{approx}{\rightsquigarrow} t(n-1)$$

en vertu du théorème central limite et de la convergence de S^2 vers σ^2. Nous en avons déduit que l'intervalle de confiance classique (à 95 %) propre au cas

d'une loi mère gaussienne :

$$IC_{0,95}(\mu) = [\overline{x} - t_{0,975}^{(n-1)} \frac{s}{\sqrt{n}} \ , \ \overline{x} + t_{0,975}^{(n-1)} \frac{s}{\sqrt{n}}]$$

fournit une bonne approximation pour une loi quelconque.

Influence de valeurs extrêmes ou aberrantes

Il a été dit qu'en principe $n \geq 30$ suffit. Cependant l'approximation par une loi de Student posera problème pour des v.a. dont les queues de distribution sont allongées et peuvent produire des observations très éloignées du centre. Si l'on étudie, par exemple, le niveau maximum annuel de crue d'une rivière sur les cent dernières années celui-ci reste en général assez semblable mais on trouvera quelques cas de valeurs exceptionnelles. Une valeur très excentrée va influencer fortement la valeur de $\overline{x}$ et encore plus, car interviennent des écarts au carré, celle de la variance s^2, rendant ainsi ces statistiques trop instables pour garantir l'approximation par une loi de Student si la taille de l'échantillon n'est pas très élevée. Pour les mêmes raisons, si les observations sont contaminées par des *valeurs aberrantes* l'approximation sera défaillante. Ceci peut provenir, par exemple, d'erreurs dans le recueil des informations ou de présences de valeurs étrangères au phénomène étudié (dans les sondages, présence d'individus distincts n'appartenant pas à la population). Si l'on soupçonne la présence de valeurs très extrêmes ou aberrantes on peut soit éliminer purement et simplement les valeurs trop éloignées par examen de la distribution des observations (histogramme), soit réduire leurs poids dans le calcul de la moyenne et de la variance. On définit ainsi des *M-estimateurs* dont l'étude des propriétés fait l'objet de la *théorie de la robustesse*. En particulier la *moyenne α-tronquée*, appropriée si la distribution est à peu près symétrique, est un M-estimateur facile à mettre en oeuvre : elle consiste à rejeter un pourcentage d'observations égal à $100(\frac{\alpha}{2})$ sur chaque extrémité. On peut également renoncer à la moyenne comme valeur caractéristique de position centrale de la distribution et préférer la médiane qui ne présente pas les mêmes inconvénients.

8.2.2 Estimation de la variance σ^2

On privilégiera l'estimateur S^2 dont on sait qu'il est sans biais et convergent (voir section 6.5.3). Pour ce qui concerne un intervalle de confiance la procédure classique obtenue dans le cas gaussien (voir section 7.4.2) ne peut être utilisée car $(n-1)S^2/\sigma^2$ ne suit plus une loi $\chi^2(n-1)$ dès que l'on s'écarte de cette hypothèse, y compris quand n tend vers l'infini. On a vu toutefois (note 5.5) que le théorème central limite s'applique à S^2, moyennant l'existence du moment d'ordre 4, et l'on peut donc établir la loi asymptotique de S^2 dans le cas général sachant que (voir exercice 5.4) :

$$V(S^2) = \frac{1}{n}(\mu'_4 - \frac{n-3}{n-1}\sigma^4) \ ,$$

où μ'_4 est le moment centré d'ordre 4 de la loi mère. Comme $V(S^2)$ est asymptotiquement équivalent à $\frac{1}{n}(\mu'_4 - \sigma^4)$ on a, après centrage et réduction, le résultat suivant :

$$\frac{\sqrt{n}(S^2 - \sigma^2)}{\sqrt{\mu'_4 - \sigma^4}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1)\ .$$

En écrivant cette statistique centrée et réduite sous la forme :

$$\frac{\sqrt{n}(S^2 - \sigma^2)}{\sigma^2\sqrt{\mu'_4/\sigma^4 - 1}}$$

on fait apparaître le *coefficient de curtose* μ'_4/σ^4 qui vaut 3 pour la loi de Gauss, est supérieur à 3 pour une loi à pic plus prononcé au mode et queues plus allongées, est inférieur à 3 pour une loi à pic plus plat et queues courtes. Dans le cas gaussien cette expression est bien, au facteur $\sqrt{(n-1)/n}$ près, la version centrée et réduite de $(n-1)S^2/\sigma^2$ puisque la loi $\chi^2(n-1)$ est de moyenne $n-1$ et de variance $2(n-1)$. Par ailleurs on a vu (voir la remarque en section 5.8) que la loi du khi-deux tend à devenir gaussienne quand n tend vers l'infini.

Pour construire un intervalle de confiance asymptotique, on peut envisager d'estimer $\mu'_4 - \sigma^4$ par sa version empirique $M'_4 - S^4$ (ou le coefficient de curtose par la curtose empirique M'_4/S^4), mais la convergence est lente et un nombre important d'observations sera nécessaire pour espérer une bonne approximation. On peut recourir à une approche dite par rééchantillonnage dont l'intérêt est général et c'est pourquoi nous y consacrons une section spécifique. Cette approche sera également appropriée pour l'estimation de l'écart-type.

8.3 Estimation d'un quantile

Nous supposerons que la v.a. soit continue pour que tout quantile existe. Pour qu'il y ait unicité nous supposerons aussi que le support de la densité f soit un intervalle $[a, b]$, où éventuellement $a = -\infty$ et/ou $b = +\infty$, de façon que F soit strictement croissante sur l'ensemble des valeurs de x telles que $0 < F(x) < 1$ (voir section 1.5). Nous considérerons essentiellement la médiane, les développements étant similaires pour un quantile quelconque.

Notons $\widetilde{\mu}$ la médiane de la loi mère et $\widetilde{X}$ la médiane empirique définie par :

$$\widetilde{X} = \begin{cases} X_{(m)} & \text{si } n = 2m-1 \\ \frac{1}{2}(X_{(m)} + X_{(m+1)}) & \text{si } n = 2m \end{cases},$$

où $X_{(m)}$ est la statistique d'ordre m (voir section 5.6). $\widetilde{X}$ est l'estimateur naturel de $\widetilde{\mu}$. Dans la proposition 5.12 nous avons donné la loi d'une statistique d'ordre quelconque, ce qui s'applique directement au cas de n impair. Pour n

fini $\widetilde{X}$ n'est pas nécessairement sans biais et sa variance n'a pas d'expression simple. On doit se contenter de la propriété asymptotique suivante, que nous admettrons.

Proposition 8.1 *Soit une loi continue de densité f et de médiane $\widetilde{\mu}$, et soit $\widetilde{X}_n$ la médiane d'un n-échantillon $X_1, X_2, \cdots, X_n$ issu de cette loi. On a :*

$$nV(\widetilde{X}) \underset{n\to\infty}{\longrightarrow} \frac{1}{4[f(\widetilde{\mu})]^2}$$

$$et \quad 2f(\widetilde{\mu})\sqrt{n}(\widetilde{X} - \widetilde{\mu}) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0\,;1)\ .$$

La médiane empirique est donc asymptotiquement sans biais et converge en moyenne quadratique vers $\widetilde{\mu}$ puisque sa variance tend vers 0 (on peut aussi montrer qu'elle converge presque sûrement). Étant donné que $f(\widetilde{\mu})$ est inconnu, il faudrait une estimation de cette valeur pour construire un intervalle de confiance approximatif. On dispose cependant d'une approche directe pour n fini.

Soit $\widetilde{N}$ le nombre d'observations inférieures ou égales à $\widetilde{\mu}$. Pour chaque X_i la probabilité d'être inférieur ou égal à $\widetilde{\mu}$ vaut $\frac{1}{2}$, et $\widetilde{N} \rightsquigarrow \mathcal{B}(n, \frac{1}{2})$. On a donc :

$$P(l_1 \leq \widetilde{N} \leq l_2) = \sum_{k=l_1}^{l_2} \binom{n}{k} \frac{1}{2^n}.$$

Choisissons l_1 et l_2 tels que, d'une part, la probabilité ci-dessus soit supérieure ou égale à 0,95 et au plus proche de cette valeur et, d'autre part, que l_2 soit égal à $n-\ l_1$ ou le plus proche possible pour avoir l'intervalle $[l_1, l_2]$ le plus symétrique possible par rapport à $n/2$ et donc le plus étroit. Notons que l'événement $(X_{(l_1)} \leq \widetilde{\mu})$ signifie qu'il y a au moins l_1 observations inférieures ou égales à $\widetilde{\mu}$ et il est donc identique à $(l_1 \leq \widetilde{N})$ et, de même, l'événement $(\widetilde{\mu} < X_{(l_2+1)})$ est identique à l'événement $(\widetilde{N} \leq l_2)$. La probabilité ci-dessus est donc égale à $P(X_{(l_1)} \leq \widetilde{\mu} < X_{(l_2+1)})$ et ceci quel que soit $\widetilde{\mu}$. Pour un échantillon réalisé $x_1, x_2, \ldots, x_n$, ceci fournit donc un intervalle de confiance à 95 % pour $\widetilde{\mu}$:

$$IC_{0,95}(\widetilde{\mu}) = [x_{(l_1)}, x_{(l_2+1)}[\,.$$

Pratiquement on peut fermer l'intervalle à droite, les statistiques d'ordre étant des v.a. continues.

Exemple 8.1 Soit $n = 20$. On a pour $\widetilde{N} \rightsquigarrow \mathcal{B}(20;0{,}5)$: $P(\widetilde{N} \leq 5) = 0{,}021$ (mais $P(\widetilde{N} \leq 6) = 0{,}58$) et par symétrie $P(\widetilde{N} \geq 15) = 0{,}021$ d'où $P(6 \leq \widetilde{N} \leq 14) = 0{,}958$ et $IC_{0,95}(\widetilde{\mu}) = [x_{(6)}, x_{(15)}]$.

On pourra utiliser assez vite l'approximation gaussienne du fait que $p = 1/2$ (le critère $np \geq 5$ et $n(1-p) \geq 5$ de la section 5.8.3 revenant à $n \geq 10$). Avec correction de continuité on obtient ici 0,022 pour $P(\widetilde{N} \leq 5)$. ■

Pour le quantile x_q d'ordre q considérons $X_{[nq]+1}$où $[nq]$ est la partie entière de nq. La proposition s'applique à $X_{[nq]+1}$ avec :

$$nV(X_{[nq]+1}) \xrightarrow[n\to\infty]{} \frac{q(1-q)}{[f(x_q)]^2}$$

$$\text{et} \quad \frac{f(x_q)}{\sqrt{q(1-q)}}\sqrt{n}(X_{[nq]+1} - x_q) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,;1)\,.$$

L'intervalle de confiance pour x_q est obtenu de façon analogue à partir de la loi $\mathcal{B}(n, q)$.

Note 8.1 De la proposition 8.1 on peut déduire la précision relative asymptotique de la médiane empirique par rapport à la moyenne empirique dans le cas gaussien, par exemple. Pour cette loi $f(\widetilde{\mu}) = f(\mu) = 1/\sqrt{2\pi}\sigma$ et $V(\widetilde{X})$ est donc équivalent à $\frac{\pi}{2}\frac{\sigma^2}{n}$ quand n tend vers l'infini. La variance asymptotique de la médiane est donc $\pi/2$ fois plus grande.

8.4 Les méthodes de rééchantillonnage

8.4.1 Introduction

Ces méthodes ont pour principe de simuler la variabilité des estimateurs en tirant des échantillons à l'intérieur de l'échantillon recueilli. La méthode du *jackknife* (littéralement couteau de poche ou canif) est la plus ancienne et effectue des tirages déterministes. La méthode du *bootstrap* (chausse-pied) constitue une généralisation du jackknife qui n'a pu être conçue que dès lors que de puissants moyens informatiques étaient disponibles. Du fait qu'elle effectue des tirages au hasard elle est de portée beaucoup plus générale. Le jackknife a été développé initialement par M. Quenouille et J. Tukey dans les années 1950 pour réduire le biais d'un estimateur donné, puis a été envisagé pour obtenir des intervalles de confiance. Le bootstrap a été proposé par Efron (1979). Notons que ces méthodes s'appliquent aussi bien dans le cas discret que dans le cas continu.

Les estimateurs non paramétriques du maximum de vraisemblance

Ces estimateurs sont les estimateurs de référence du jackknife comme du bootstrap et il est utile de les expliciter. Pour une caractéristique ω de la loi définie par sa fonction de répartition F (en bref nous dirons la loi F) son estimateur du maximum de vraisemblance $\widehat{\omega}$ est obtenu par sa version empirique, c'est-à-dire la caractéristique de même nature calculée sur l'échantillon : la moyenne de la loi est estimée par la moyenne de l'échantillon $\overline{X}$, la variance par $\widetilde{S}^2$, la médiane de la loi par la médiane de l'échantillon, etc. Ceci découle du fait que la fonction de répartition empirique F_n (définie en section 5.7) est l'*estimateur fonctionnel* (i.e. une fonction pour estimer une fonction) du maximum de vraisemblance pour F (voir plus loin, section 8.5.3).

Notons que le passage d'une caractéristique théorique à la caractéristique empirique correspondante est immédiat dans le cas d'une loi discrète (et en particulier dans le cas d'un tirage au hasard dans une population finie) car il suffit d'appliquer à l'échantillon la même formule de définition que celle de la caractéristique. Si la loi est continue et que la caractéristique est propre à ce type de loi, le passage peut être plus délicat du fait que F_n n'est pas continue. C'est par exemple le cas d'un mode de la loi (maximum de la dérivée de F) qui n'a pas d'équivalent direct sur l'échantillon. On pourra alors utiliser un lissage de F_n en continu du type de celui présenté en section 8.5.3.

Note 8.2 Puisque la loi est parfaitement définie à partir de F, une caractéristique ω de la loi F peut s'exprimer comme une application $\omega(F)$qui à une fonction fait correspondre un réel. On dira que ω est un opérateur fonctionnel (en bref une fonctionnelle). Par exemple la moyenne μ est égale à l'intégrale de Riemann-Stieltjes définie dans la note 2.1 : $\mu(F) = \int_{\mathbb{R}} x dF(x)$ et la moyenne empirique est la statistique fonctionnelle correspondante obtenue en remplaçant F par la fonction de répartition empirique F_n : $\overline{x} = \int_{\mathbb{R}} x dF_n(x)$ (voir note 5.4). Plus généralement, ce que nous avons appelé la version empirique de $\omega(F)$ est $\omega(F_n)$. Comme la fonction de répartition empirique F_n est l'estimation du maximum de vraisemblance de F (voir section 8.5.3), par voie de conséquence $\omega(F_n)$, en tant que fonction de F_n, est l'estimation du maximum de vraisemblance de $\omega(F)$.

8.4.2 La méthode du jackknife

Nous illustrons d'abord le principe du jackknife pour l'estimation de l'écart-type σ de la loi. Soit l'estimateur de référence :

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

qui est généralement biaisé (voir exercice 6.4 pour le cas de la loi de Gauss) et soit s l'estimation correspondante pour une réalisation $x_1, x_2, \cdots, x_n$ de l'échantillon. L'estimation du jackknife est obtenue de la façon suivante.

On calcule la valeur, notée s_{-1}, de l'écart-type du sous-échantillon obtenu en **omettant la valeur** x_1 :

$$s_{-1} = \sqrt{\frac{1}{n-2}\sum_{i=2}^{n}(x_i - \overline{x})^2},$$

puis la valeur $s_{*1} = ns - (n-1)s_{-1}$. On répète cette opération en omettant à tour de rôle chacune des observations pour obtenir n *pseudo-valeurs*

$s_{*1}, s_{*2}, \cdots, s_{*n}$ avec, donc :

$$s_{-i} = \sqrt{\frac{1}{n-2} \sum_{j=1, j\neq i}^{n} (x_j - \overline{x})^2}$$
$$s_{*i} = ns - (n-1)s_{-i} \ .$$

L'estimation du jackknife est alors la moyenne des pseudo-valeurs, notée $\overline{s}_*$. Un intervalle de confiance approché peut être obtenu en appliquant à la série des n pseudo-valeurs le résultat de la section 6.4.1 concernant la moyenne d'un échantillon aléatoire gaussien. Ainsi on calcule la variance des pseudo-valeurs :

$$s^2_{JK} = \frac{1}{n-1} \sum_{i=1}^{n} (s_{*i} - \overline{s}_*)^2,$$

d'où :

$$IC_{0,95}(\sigma) \simeq [\overline{s}_* - t^{(n-1)}_{0,975} \frac{s_{JK}}{\sqrt{n}} \ , \ \overline{s}_* + t^{(n-1)}_{0,975} \frac{s_{JK}}{\sqrt{n}}] \ .$$

De façon générale soit ω une caractéristique de la loi et T_n un estimateur convergent de ω, typiquement l'estimateur du maximum de vraisemblance. Soit T_n^{-i} l'estimateur calculé en omettant X_i. On définit les pseudo-valeurs :

$$T_n^{*i} = nT_n - (n-1)T_n^{-i} \ , \ i = 1, \ldots, n.$$

L'estimateur du jackknife fondé sur T_n est alors $T_n^* = \frac{1}{n} \sum_{i=1}^{n} T_n^{*i}$.

Comme il a été dit en introduction cet estimateur a été proposé à l'origine pour réduire le biais éventuel de T_n, en vertu du résultat suivant.

Proposition 8.2 *Si le biais de T_n est de la forme $\frac{c}{n}$, où c est une constante, alors T_n^*, l'estimateur du jackknife fondé sur T_n, est sans biais.*

En effet, comme $E(T_n) = \omega + \frac{c}{n}$, on a, pour tout i, $E(T_n^{-i}) = \omega + \frac{c}{n-1}$ puisque T_n^{-i} est le même estimateur appliqué au $(n-1)-$échantillon aléatoire $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$. Ainsi :

$$\begin{aligned} E(T_n^{*i}) &= nE(T_n) - (n-1)E(T_n^{-i}) \\ &= n\omega + c - (n-1)[\omega + \frac{c}{n-1}] \\ &= \omega \ , \end{aligned}$$

d'où $E(T_n^*) = \frac{1}{n} \sum_{i=1}^{n} E(T_n^{*i}) = \omega$.

Si le biais est de la forme $\frac{c_1}{n} + \frac{c_2}{n^2} + \frac{c_3}{n^3} + \cdots$ on montre aisément de la même façon que le premier terme disparaît dans le biais de T_n^*. Par conséquent, au

moins pour des situations de ce type, il y a réduction de biais. Si l'on applique, par exemple, la procédure du jackknife à la variance empirique $\widetilde{S}_n^2$ dont le biais pour estimer σ^2 est $-\frac{1}{n}\sigma^2$ (voir proposition 5.2), on trouve que l'estimateur du jackknife est la variance de l'échantillon S_n^2 qui est sans biais (voir exercices). Notons incidemment que pour l'estimateur $\overline{X}_n$ de la moyenne μ qui est sans biais, l'estimateur du jackknife est $\overline{X}_n$ lui-même.

En général, la loi étant totalement inconnue, on ne connaît pas la forme du biais (comme par exemple pour l'écart-type S dans l'illustration ci-dessus), mais on s'attend à ce qu'il soit de toute façon réduit par la procédure décrite.

Outre la réduction du biais, l'intérêt du jackknife, primordial ici, est de permettre l'estimation de l'écart-type de T_n et la possibilité de construire un intervalle de confiance approché. La proposition qui suit va nous y conduire.

Proposition 8.3 *Soit T_n^* l'estimateur du jackknife de la caractéristique ω, reposant sur un estimateur convergent T_n, et soit $S_{n,JK}^2$ la variance des pseudo-valeurs. Alors, sous certaines conditions concernant la forme de la statistique :*

$$\frac{T_n^* - \omega}{S_{n,JK}/\sqrt{n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0\,;1)\ .$$

Nous admettrons cette proposition. Elle résulte du fait que les pseudo-valeurs tendent à être indépendantes et gaussiennes pour une grande variété de statistiques. En appliquant l'intervalle de confiance de la section 7.4.1 pour la moyenne de v.a. i.i.d. gaussiennes, on déduit l'intervalle de confiance approché pour ω :

$$IC_{0,975}(\omega) \simeq [t_n^* - t_{0,975}^{(n-1)} \frac{s_{n,JK}}{\sqrt{n}}\ ,\ t_n^* + t_{0,975}^{(n-1)} \frac{s_{n,JK}}{\sqrt{n}}]\ ,$$

où t_n^* et $s_{n,JK}$ sont les réalisations respectives de T_n^* et de $S_{n,JK}$.

Ceci s'applique, en particulier, à l'écart-type comme nous l'avons vu plus haut (voir une application dans les exercices) et également pour estimer la variance σ^2. Dans ce dernier cas, en prenant l'estimateur du jackknife reposant sur la variance empirique $\widetilde{S}^2$, on établit que la variance des pseudo-valeurs $S_{n,JK}^2$ est égale à :

$$\frac{n^3}{(n-1)(n-2)^2}(M_4' - \widetilde{S}^4)\ ,$$

ce qui conduit à une procédure d'intervalle de confiance très proche (et asymptotiquement équivalente) de l'approche asymptotique proposée en section 8.2.2. En effet, dans cette approche, on trouvait simplement $M_4' - S^4$ en lieu et place de l'expression ci-dessus.

Les conditions de validité de la proposition ne sont pas simples à expliciter. Si la statistique est de la forme $\frac{1}{n}\sum_{i=1}^n g(X_i)$ où g est une fonction quelconque,

alors la proposition est vérifiée. C'est le cas de tous les moments simples. Si la forme est proche cela reste vrai, comme par exemple pour les moments centrés, en particulier pour la variance empirique, et aussi pour l'écart-type. En revanche, la médiane qui, dans sa version théorique, s'exprime par $F^{-1}(\frac{1}{2})$ a une forme très éloignée. Le jackknife est alors inadapté car $S_{n,JK}$ ne converge pas vers la valeur de l'écart-type de la médiane. Il en va de même pour d'autres statistiques fonctions des statistiques d'ordres : quantiles, étendue $X_{(n)} - X_{(1)}$, distance interquartiles.

Note 8.3 Pour préciser quelque peu le domaine de validité du jackknife exprimons une caractéristique ω comme une expression fonctionnelle $\omega(F)$. Une fonctionnelle est dite linéaire si $\omega(a_1F_1 + a_2F_2) = a_1\omega(F_1) + a_2\omega(F_2)$. Dans ce cas on montre que $\omega(F)$ est de la forme :

$$\omega(F) = \int_{\mathbb{R}} g(x)dF(x) = E_F(g(X)).$$

Pour la statistique correspondante $\omega(F_n)$ du maximum de vraisemblance cela se traduit par $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$. Ceci est évidemment le cas de la moyenne empirique et de tout autre moment empirique non centré.

Un moment centré n'est pas strictement de cette forme. Par exemple la variance $E_F([X - E_F(X)]^2)$ est l'espérance d'une fonction qui dépend elle-même de F. La condition pour que le jackknife soit opérant au niveau de la convergence de $S_{n,JK}$ est que la caractéristique, et donc la statistique du MV, soit une fonctionnelle linéaire ou pouvant être raisonnablement approchée par une fonctionnelle linéaire. Ceci est réalisable pour la variance empirique (qui est une fonctionnelle quadratique) mais pas dans le cas de la médiane qui est trop fortement non linéaire.

Le jackknife peut être utilisé pour des couples (et des n-uplets) de v.a., par exemple pour la corrélation entre deux variables, pour la moyenne du ratio de deux variables. Il s'étend également à des situations autres que des échantillons aléatoires simples. Par ailleurs, différentes variantes du jackknife initial ont été proposées. En particulier, pour de très grands échantillons, il est pratiquement aussi efficace de l'appliquer en omettant non pas chaque observation mais des groupes de k observations, ceci afin d'accélérer les calculs. Dans le cas de la médiane le fait de grouper les observations avec k de l'ordre de $\sqrt{n}$ permet même d'assurer la convergence selon la proposition 8.3 et donc d'appliquer l'intervalle de confiance qui en découle.

Rien ne s'oppose à ce qu'on utilise cette méthode dans un cadre paramétrique pour des fonctions du paramètre complexes. Par exemple on pourra estimer $e^{-\lambda}$, la probabilité qu'il n'y ait aucune occurrence dans une unité de temps pour une loi de Poisson, en se fondant sur l'estimateur du maximum de vraisemblance $e^{-\overline{X}}$ (ceci est à rapprocher de l'exemple 6.13).

En ce qui concerne l'approximation asymptotique de l'intervalle de confiance issu du jackknife il est difficile de savoir à partir de quelle taille d'échantillon

elle devient satisfaisante. Pour les petits échantillons le bootstrap offre une alternative plus sûre.

8.4.3 La méthode du bootstrap

Le bootstrap est une approche très générale pour des situations les plus variées. Il est certes plus coûteux en calcul que le jackknife mais donne en général des estimateurs de variance plus faible. On a pu le voir comme une généralisation du jackknife ou, plus exactement, le jackknife a pu être considéré comme une forme appauvrie du bootstrap. Au lieu de rééchantillonner au hasard un grand nombre de fois, comme la théorie montre qu'il convient de le faire et comme le fait le bootstrap, le jackknife se contente de choisir des échantillons bien déterminés en nombre limité à n.

Nous indiquerons sans démonstration[1] comment le bootstrap permet d'estimer la variance (donc la précision) d'un estimateur, ce qui débouche sur la construction d'un intervalle de confiance approché. Pour ne pas alourdir les notations on passera indifféremment d'une variable aléatoire à sa réalisation, le contexte permettant de reconnaître l'une ou l'autre de ces entités.

Soit $\widehat{\omega}$ un estimateur d'une caractéristique ω de la loi mère, cette loi étant quelconque, continue ou discrète, de fonction de répartition inconnue F. On s'intéresse à la variance $V_F(\widehat{\omega})$ de $\widehat{\omega}$ ou, plutôt, à une estimation de cette variance puisque F est inconnue. Pour aboutir à cette estimation on opère selon les étapes suivantes :

1. soit $x_1, x_2, \ldots, x_n$ l'échantillon réalisé. On effectue n tirages au hasard **avec remise** parmi les valeurs $x_1, x_2, \ldots, x_n$ (on s'attend à des répétitions car il est très improbable de tirer les n valeurs distinctes initiales). On calcule l'estimation $\widehat{\omega}^*$ obtenue sur la base de ce nouvel échantillon.
2. on répète l'opération précédente M fois pour obtenir une série d'estimations $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_M^*$.
3. l'estimation de la variance propre à $\widehat{\omega}$ est fournie par la variance descriptive de cette série d'observations, i.e. :

$$s^{2*}(\widehat{\omega}) = \frac{1}{M-1} \sum_{k=1}^{M} (\widehat{\omega}_k^* - \overline{\widehat{\omega}}^*)^2$$

où $\overline{\widehat{\omega}}^*$ désigne la moyenne de la série.

On montre que lorsque M tend vers l'infini, l'estimateur issu de cette procédure tend presque sûrement vers l'estimateur du maximum de vraisemblance de $V_F(\widehat{\omega})$. En pratique $M = 100$ fournit une approximation suffisante de cet EMV car l'écart sera alors négligeable par rapport à l'erreur d'estimation du maximum de vraisemblance lui-même.

[1] Pour de plus amples développements on pourra consulter les ouvrages de référence indiqués en fin de section.

Note 8.4 Dénotons $V_F(\widehat{\omega})$ comme une forme fonctionnelle $\sigma^2_{\widehat{\omega}}(F)$ pour faire apparaître le lien avec la loi mère de l'échantillon. Sauf dans les cas simples cette fonctionnelle ne peut être explicitée ce qui n'a aucune importance ici. Prenons toutefois comme illustration élémentaire le cas où $\widehat{\omega}$ est $\overline{X}$, l'estimateur «moyenne de l'échantillon». On sait que sa variance est la variance de la loi mère divisée par n et l'on peut donc expliciter le lien :

$$\sigma^2_{\overline{X}}(F) = \frac{1}{n}\int_{\mathbb{R}}(x-\mu)^2 dF(x)$$

où μ est en fait $\mu(F) = \int_{\mathbb{R}} x dF(x)$. Si, comme dans ce cas, la forme fonctionnelle $\sigma^2_{\widehat{\omega}}(F)$ est connue on peut l'estimer par sa version empirique $\sigma^2_{\widehat{\omega}}(F_n)$ obtenue en remplaçant F par F_n, qui est l'estimateur du maximum de vraisemblance de $\sigma^2_{\widehat{\omega}}(F)$ (voir note 8.2). Ainsi pour la moyenne, $\sigma^2_{\overline{X}}(F_n)$ devient $\widetilde{S}^2/n$. Mais si $\widehat{\omega}$ est, par exemple, l'estimateur «médiane de l'échantillon» on ne connaît pas l'expression de sa variance. C'est alors que la procédure bootstrap devient particulièrement précieuse.

Si l'on connaissait F on pourrait estimer $\sigma^2_{\widehat{\omega}}(F)$ avec toute la précision souhaitée par simulation : on sait générer des échantillons de taille n issus de la loi F (méthode de Monte-Carlo), sur chaque échantillon on calculerait la statistique $\widehat{\omega}$ et, pour un grand nombre d'échantillons ainsi générés, on approcherait la distribution réelle de $\widehat{\omega}$. En calculant, par exemple, la variance empirique des valeurs de $\widehat{\omega}$ ainsi générées on obtiendrait une estimation de la vraie variance de $\widehat{\omega}$, d'autant plus précise que le nombre d'échantillons générés serait grand, en vertu de la loi des grands nombres. Le processus est identique lorsqu'il s'agit d'estimer $\sigma^2_{\widehat{\omega}}(F_n)$ en considérant maintenant l'échantillon réalisé $x_1, x_2, \ldots, x_n$ comme une population en soi, la distribution de celle-ci étant caractérisée par F_n. Si l'on parcourt l'univers de tous les échantillons possibles ($M \to \infty$) et que l'on examine comment varie la statistique $\widehat{\omega}$ en calculant sa variance on obtient $\sigma^2_{\widehat{\omega}}(F_n)$. Par exemple la variance des moyennes des échantillons convergera vers $\widetilde{S}^2/n$. Pour cette raison $\sigma^2_{\widehat{\omega}}(F_n)$ est appelé **estimation du bootstrap** de la variance de $\widehat{\omega}$ puisque la procédure de rééchantillonnage permet de se rapprocher autant que l'on veut de l'EMV. Notons au passage que le jackknife ne parcourt que n échantillons particuliers ce qui explique ses performances moindres.

Ayant estimé la variance de $\widehat{\omega}$ on peut obtenir un intervalle de confiance approché pour ω en supposant que $\widehat{\omega}$ soit asymptotiquement sans biais et gaussien :

$$IC_{0,95}(\omega) \simeq [\widehat{\omega} - 1,96\, s^*(\widehat{\omega})\,,\ \widehat{\omega} + 1,96\, s^*(\widehat{\omega})\,]\,.$$

L'approximation gaussienne est fréquemment légitime, en particulier si $\widehat{\omega}$ est lui-même une estimation du maximum de vraisemblance de ω, la normalité asymptotique de l'EMV énoncée en proposition 6.11 dépassant le strict cadre paramétrique. Mais celle-ci n'est pas assurée pour tous les types de statistiques ou bien elle peut être trop lente pour fournir une approximation satisfaisante au vu de la taille n de l'échantillon. En fait on peut contrôler l'hypothèse de

normalité en examinant l'histogramme des valeurs $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_M^*$ car il reflète la distribution de $\widehat{\omega}$ (voir section 8.5.2).

On peut améliorer cette méthode «basique» par la *méthode studentisée*. Cette méthode s'applique toutefois si $\widehat{\omega}$ tend à être gaussien quand $n \to \infty$ avec une variance de forme équivalente à σ_ω^2/n où σ_ω^2 est une constante (par exemple si $\widehat{\omega}$ est EMV ce peut être $1/I(\omega)$) et si l'on dispose d'un estimateur convergent de σ_ω^2 que nous noterons simplement s^2. A l'étape 1 on calcule, en plus de $\widehat{\omega}^*$, la valeur s^{2*} pour les n valeurs rééchantillonnées. L'étape 2 produit, en plus de la série des $\widehat{\omega}_k^*$ précédents, les s_k^{2*} correspondants. On calcule alors la série des valeurs $T_1^*, T_2^*, \ldots, T_M^*$ définie par :

$$T_k^* = \frac{\widehat{\omega}_k^* - \widehat{\omega}}{s_k^*/\sqrt{n}}$$

et l'on détermine les quantiles empiriques $t_{0,025}^*$ et $t_{0,975}^*$ de cette série (moyennant M grand, disons $M \simeq 1000$, pour évaluer avec suffisamment de précision de petites probabilités sur les extrémités de la distribution). Alors l'IC approché est :

$$IC_{0,95}(\omega) \simeq [\widehat{\omega} + t_{0,025}^* \ \frac{s}{\sqrt{n}} \, , \, \widehat{\omega} + t_{0,975}^* \ \frac{s}{\sqrt{n}}] \, .$$

La *méthode des percentiles* constitue une autre approche simple à mettre en oeuvre et, de ce fait, assez répandue. Elle s'applique en prenant directement pour bornes, avec $M \simeq 1000$, les quantiles empiriques $\widehat{\omega}_{0,025}^*$ et $\widehat{\omega}_{0,975}^*$ d'ordres respectifs 0,025 et 0,975 de la série $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_M^*$, soit :

$$IC_{0,95}(\omega) \simeq [\widehat{\omega}_{0,025}^* \, , \, \widehat{\omega}_{0,975}^*] \, .$$

Cette méthode présentée à l'origine comme méthode de référence des IC bootstrap est moins précise que la précédente et ne donne de bons résultats qu'à condition qu'il existe une fonction croissante h telle que $h(\widehat{\omega})$ ait une loi symétrique autour de $h(\omega)$. Cette condition est forte et évidemment invérifiable dans les situations pratiques complexes où le bootstrap est le principal recours. Cette méthode n'est donc pas sans risque.

Nous avons concentré notre attention sur l'obtention d'un intervalle de confiance dans le cadre non paramétrique. Cependant l'intérêt des valeurs bootstrap $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_M^*$ ne se limite pas à cela. En effet pour des types très divers de statistiques, il a été démontré que la distribution générée par les valeurs $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_M^*$ en faisant tendre M vers l'infini, appelée *distribution bootstrap* de la statistique $\widehat{\omega}$, converge vers la vraie distribution de $\widehat{\omega}$ quand n tend vers l'infini et ceci de façon rapide. En d'autres termes, pour une valeur de M raisonnable, la série des valeurs $\widehat{\omega}_1^*, \widehat{\omega}_2^*, \ldots, \widehat{\omega}_M^*$ reflète bien la distribution d'échantillonnage de $\widehat{\omega}$. Il faut bien distinguer la convergence pour $M \to \infty$ qui,

à un premier niveau, assure une bonne approche de l'estimation du maximum de vraisemblance de toute statistique **propre à l'échantillon réalisé** à partir de sa valeur dans la distribution bootstrap, de la convergence pour $n \to \infty$ qui, à un deuxième niveau, concerne la convergence de l'estimateur du maximum de vraisemblance vers la vraie valeur ω, même si, en vérité, les deux niveaux sont intimement liés, et ceci de façon optimale, ce qui fait la grande force du bootstrap. Notons ici que la méthode des percentiles donne en fait un intervalle de probabilité approché pour la statistique $\widehat{\omega}$.

Ainsi, au-delà de l'écart-type, toute caractéristique de la distribution d'échantillonnage de $\widehat{\omega}$ peut être estimée par sa version empirique dans la série des valeurs bootstrap. Par exemple la moyenne de la série $\overline{\widehat{\omega}}^*$ est une estimation de la moyenne $E_F(\widehat{\omega})$ de la loi de $\widehat{\omega}$. Ceci peut déboucher sur l'estimation du biais d'une statistique. Illustrons ceci pour la moyenne α-tronquée.

En présence de valeurs extrêmes il peut être préférable (au sens de l'e.q.m.) d'utiliser une moyenne α-tronquée (voir section 8.2.1), que nous notons toujours $\widehat{\omega}$, pour estimer la moyenne μ de la loi étudiée. On peut, par le bootstrap, estimer le gain en variance en comparant la variance bootstrap $s^{2*}(\widehat{\omega})$ de cet estimateur à la variance estimée de la moyenne simple, à savoir s^2/n. Mais cette moyenne α-tronquée peut être plus ou moins fortement biaisée. On peut obtenir une estimation raisonnable de ce biais par la différence $\overline{\widehat{\omega}}^* - \overline{x}$ (signalons toutefois qu'il n'est pas pour autant judicieux de rectifier de ce biais l'estimation $\widehat{\omega}$). Il est clair que ceci ne s'applique qu'en présence de valeurs extrêmes et non de valeurs aberrantes qui ne proviendraient pas de la loi F, car la théorie repose sur une série d'observations toutes issues de cette même loi F.

L'approche bootstrap est appropriée dans des situations très complexes, où il n'y aura généralement pas d'alternative, et pas uniquement dans le cadre de l'échantillonnage aléatoire simple. C'est donc un outil extrêmement précieux qui est devenu viable avec les capacités de calcul actuelles.

Signalons en particulier qu'alors que le jackknife échoue pour obtenir un intervalle de confiance pour la médiane (ou un quantile) de la loi mère, le bootstrap donne un résultat très proche de l'intervalle proposé par approche directe en section 8.3. Il est intéressant aussi de voir qu'il peut s'appliquer dans un cadre paramétrique si l'on estime une fonction du paramètre avec un estimateur dont on ignore la loi ou l'expression de la variance. On aura alors avantage à effectuer les tirages de l'étape 2 à partir de la fonction de répartition estimée $F(x;\widehat{\theta}^{MV})$ obtenue en remplaçant θ par son estimation du maximum de vraisemblance dans l'expression $F(x;\theta)$. On montre que pour M tendant vers l'infini et sous certaines conditions de régularité, l'intervalle de confiance de la méthode des percentiles tend vers celui obtenu par l'approximation normale asymptotique du MV vue en section 7.3.

Enfin il existe des variantes visant à améliorer encore le bootstrap basique : méthode des percentiles à biais corrigé, bootstrap lissé.

Pour approfondir la méthode du bootstrap on pourra consulter l'ouvrage de Davison et Hinkley (1997) pour les aspects méthodologiques et celui de Shao et Tu (1995) pour les démonstrations mathématiques.

8.5 Estimation fonctionnelle

8.5.1 Introduction

Dans le cadre non paramétrique il est pertinent de vouloir estimer, dans sa globalité, la fonction de répartition F, ou la fonction de densité f, ou la fonction de probabilité p, alors que, dans le cadre paramétrique, ces fonctions découlent du seul choix du paramètre inconnu θ. Pour les v.a. continues nous ferons l'hypothèse que F et f sont quelconques mais dérivables, donc lisses en pratique, au moins jusqu'à l'ordre 2 et parfois plus. Pour les v.a. discrètes on pourrait envisager de travailler avec de telles hypothèses pour F et p étendues à tout $\mathbb{R}$ (comme c'est le cas pour la plupart des lois discrètes usuelles), mais ceci n'est pratiquement jamais effectué car on se contente, l'ensemble des valeurs possibles $\{x_i\}$ étant connu, d'estimer $p(x_i)$ (respectivement $F(x_i)$) par les fréquences relatives (respectivement les fréquences relatives cumulées) observées. **Cette section ne concerne donc que les variables aléatoires continues.** Au sens strict on exclut donc l'étude d'une population réelle qui ne prend qu'un nombre fini de valeurs, mais on peut supposer qu'en amont d'une telle population il existe un modèle virtuel, parfois appelé modèle de *superpopulation,* qui est celui que l'on cherche à estimer.

On peut se poser la question de savoir quel est l'intérêt réel de l'estimation de f ou de F. L'intérêt le plus immédiat est de visualiser la distribution des valeurs, ce qui est propre à la fonction de densité plutôt qu'à la fonction de répartition. En effet cette dernière met mal en évidence les zones de fortes ou faibles probabilités. L'estimation de la densité peut aussi viser à une première prise de connaissance du phénomène de façon à orienter la recherche d'un modèle adéquat dans la panoplie des modèles paramétriques disponibles. Dans certaines applications techniques la densité ou la fonction de probabilité peut avoir un intérêt en soi, par exemple pour effectuer des simulations fines de processus ou pour montrer des caractéristiques très spécifiques (par exemple des points d'inflexion) ayant une interprétation physique. Quant aux caractéristiques usuelles (moyenne, variance, moments, quantiles, etc.) il arrive qu'en les estimant par les caractéristiques correspondantes de l'estimation de la fonction F ou f on améliore les méthodes d'estimation directes, mais nous n'envisagerons pas ces possibilités encore peu explorées, considérant que globalement les estimations directes vues aux sections précédentes restent préférables. Nous n'avons pas abordé plus haut l'estimation du ou des modes de la distribution (i.e. les positions des maxima de la densité) : l'estimation de la densité fournira une façon pertinente d'estimer ces caractéristiques.

Le développement de l'estimation fonctionnelle est relativement récent, notamment parce que les méthodes mobilisent de gros moyens de calcul et qu'elles s'appliquent à des échantillons de taille plutôt grande. Dans les années 1950 on s'est d'abord intéressé à l'estimation de la densité qui présente, comme on vient de le voir, un intérêt dominant.

8.5.2 L'estimation de la densité

L'histogramme comme estimation de densité

L'histogramme dont l'origine est attribuée à John Graunt au XVII^{e} siècle répondait à l'objectif d'une représentation de la distribution de valeurs et, à ce titre, peut être considéré comme une estimation de densité avant l'heure. C'est sous cet angle que nous allons l'étudier même si son intérêt réside souvent plutôt dans l'estimation de pourcentages dans des intervalles (ou «classes») bien déterminés (par exemple les classes d'âge des statistiques officielles).

Dans sa plus grande généralité un histogramme se définit à partir d'une suite double de valeurs croissantes $\{\ldots, a_{-i}, \ldots, a_{-1}, a_0, a_1, \ldots, a_i, \ldots\}$ constituant un découpage en intervalles de la droite réelle. Soit n_k la fréquence des observations situées dans l'intervalle $]a_k, a_{k+1}]$ pour un échantillon de taille n, alors l'histogramme est la fonction constante par morceaux $\widehat{f}_n$ telle que, pour tout $k \in \mathbb{Z}$:

$$\widehat{f}_n(x) = \frac{\dfrac{n_k}{n}}{(a_{k+1} - a_k)} \quad \text{pour } x \in]a_k, a_{k+1}],$$

conduisant à la représentation graphique classique en rectangles (obtenue en délimitant verticalement les intervalles). La fréquence relative n_k/n estimant la probabilité (ou la proportion, dans une population) p_k associée à l'intervalle $]a_k, a_{k+1}]$ y est divisée par la largeur $a_{k+1} - a_k$ de cet intervalle ce qui a bien valeur de densité de probabilité au sens explicité en section 1.4.

Sauf exception on choisit une «grille» de découpage $\{a_k\}$ régulière et soit, alors, h la largeur de chaque intervalle. On a :

$$\widehat{f}_n(x) = \frac{n_k}{nh} \quad \text{pour } x \in]a_k, a_{k+1}] \ .$$

Plaçons-nous maintenant dans le cadre d'un $n-$échantillon aléatoire de loi mère de densité f continue sur tout $\mathbb{R}$ et étudions les propriétés d'échantillonnage de la v.a. (notée simplement comme précédemment) $\widehat{f}_n(x) = N_k/nh$ où N_k est le nombre aléatoire de valeurs tombant dans $]a_k, a_{k+1}]$, x étant fixé dans cet intervalle. On a $N_k \rightsquigarrow \mathcal{B}(n, p_k)$, d'où :

$$E(\widehat{f}_n(x)) = \frac{np_k}{nh} = \frac{p_k}{h}$$
$$V(\widehat{f}_n(x)) = \frac{np_k(1-p_k)}{n^2h^2} = \frac{p_k(1-p_k)}{nh^2} \ .$$

Comme $p_k = \int_{a_k}^{a_{k+1}} f(x)dx$, $\frac{p_k}{h}$ est la valeur moyenne de f sur $[a_k, a_{k+1}]$ et $\widehat{f}_n(x)$ n'est donc sans biais que pour la ou les valeurs de x dans $]a_k, a_{k+1}]$ où f prend cette valeur moyenne. Nous désignons par x_k^* l'une de ces valeurs, i.e. telle que $f(x_k^*) = \frac{1}{h}\int_{a_k}^{a_{k+1}} f(x)dx = \frac{p_k}{h}$. Pour toute valeur x où $f(x)$ diffère de $f(x_k^*)$, $\widehat{f}_n(x)$ est un estimateur biaisé de $f(x)$ dont le biais est égal à $f(x_k^*) - f(x)$.

Considérons le comportement asymptotique de $\widehat{f}_n(x)$. Quand $n \to \infty$ le biais qui ne dépend pas de n ne peut tendre vers zéro. Sauf à faire tendre simultanément la largeur d'intervalle h vers zéro, auquel cas x tend nécessairement vers x_k^* et, par continuité de f, $f(x_k^*) - f(x)$ tend donc vers zéro. Quant à la variance, du fait que $p_k(1-p_k)/h = (1-p_k)f(x_k^*)$, elle ne peut tendre vers zéro que si, simultanément, $nh \to \infty$. En d'autres termes la largeur d'intervalle doit tendre vers 0 mais de façon infiniment moins « rapide » que $1/n$, par exemple en choisissant $h = c/\sqrt{n}$. Puisque $E(N_k) = np_k = nhf(x_k^*)$ la condition $nh \to \infty$ assure que le nombre attendu de valeurs dans $]a_k, a_{k+1}]$ tende vers l'infini. Concrètement cela se traduit de la façon suivante : plus n est grand plus il y a avantage à resserrer les intervalles mais pas trop, afin de garder de grandes valeurs de n_k.

Ces conditions qui assurent la convergence en moyenne quadratique - soit $n \to \infty$, $h \to 0$, $nh \to \infty$ - restent nécessaires pour assurer d'autres modes de convergence, notamment en probabilité ou presque sûrement. De plus elles se retrouvent pour tous les types d'estimateurs fonctionnels comme nous en verrons des exemples plus loin. La proposition suivante, que nous admettrons, vient préciser la forme asymptotique de l'erreur quadratique moyenne de $\widehat{f}_n(x)$ pour x fixé.

Proposition 8.4 *(Friedman et Diaconis, 1981) En tout x où f est deux fois dérivable on a :*

$$eqm(\widehat{f}_n(x)) = \frac{h^2}{12}[f'(x)]^2 + \frac{f(x)}{nh} + o(h^2) + o(\frac{1}{nh}) .$$

Ainsi avec les conditions $n \to \infty, h \to 0, nh \to \infty$, l'e.q.m. est asymptotiquement équivalente à $\frac{h^2}{12}[f'(x)]^2 + \frac{f(x)}{nh}$, le premier terme étant dû au biais (au carré) et le deuxième à la variance. L'intérêt de ce résultat est d'établir la *vitesse de convergence* de l'e.q.m. vers zéro, dans le cas où h est choisi de façon optimale, pour la comparer plus loin avec un estimateur plus élaboré. En annulant sa dérivée par rapport à h on trouve que cette dernière expression est minimale pour :

$$h = \{\frac{6f(x)}{[f'(x)]^2}\}^{1/3} n^{-1/3}$$

et l'e.q.m. est asymptotiquement équivalente, avec cet h optimal, à $k(x)n^{-2/3}$ où $k(x)$ dépend de $f(x)$ et de $f'(x)$. On dira que la vitesse de convergence de l'e.q.m. est en $n^{-2/3}$.

Remarques

1. La construction de l'histogramme dépend de deux paramètres : h, la largeur des intervalles, et a_0, la position de l'origine de la grille. En fait le choix de a_0 n'est pas crucial (et ne subsiste pas dans l'approche plus élaborée qui va suivre) alors que celui de h est déterminant et incontournable. Notons que le résultat précédent concernant le h optimal n'est d'aucun secours en pratique car cette valeur dépend de $f(x)$ et de $f'(x)$ qui sont inconnus. Différentes règles empiriques, étrangères aux considérations asymptotiques ci-dessus, ont été proposées en statistique descriptive, par exemple : choisir un nombre d'intervalles sur l'étendue des observations égal à $\sqrt{n}$.
2. Si f est discontinue aux bornes de son support (voir la loi uniforme, la loi exponentielle ou la loi de Pareto) les résultats développés ne sont pas valables en ces bornes. De plus si celles-ci sont inconnues l'histogramme pose problème.
3. L'histogramme, c'est-à-dire la fonction $\widehat{f}_n$, est discontinu alors même que f est continue. On peut donc songer à le rendre continu pour, sans doute, améliorer son efficacité. C'est l'idée qui prévaut pour le *polygone des fréquences* (ligne brisée reliant les milieux des «plateaux» de l'histogramme) qui reste cependant peu usité. La méthode de la section suivante va proposer une solution plus performante.

Les estimateurs à noyaux

Définition L'origine de la méthode des noyaux est due à Rosenblatt (1956). Celui-ci a proposé une sorte d'histogramme mobile où la *fenêtre* de comptage des observations se déplace avec la valeur de x. La densité en x est estimée par la fréquence relative des observations dans l'intervalle $[x-h, x+h]$, donc centré sur x, divisée naturellement par la largeur de l'intervalle $2h$. On appelle h la *largeur de fenêtre* (bien que cette largeur soit en fait égale à $2h$). Pour des raisons qui apparaîtront plus loin nous écrivons l'estimation ainsi obtenue à partir des observations $x_1, x_2, \cdots, x_n$ sous la forme suivante (conservant encore la même notation $\widehat{f}_n$) :

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

$$\text{où} \quad K(u) = \frac{1}{2} \text{ si } u \in [-1, +1] \text{ et } 0 \text{ sinon.}$$

En effet $x_i \in [x-h, x+h]$ si et seulement si $\frac{x-x_i}{h} \in [-1, +1]$ et x_i est alors comptabilisé $1/2$. Ainsi $\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$ est égal au nombre d'observations tombant dans $[x-h, x+h]$ divisé par 2 pour obtenir la division de la fréquence relative par $2h$. Comme K est discontinue en ± 1, $\widehat{f}_n(x)$ présente des petits sauts de discontinuité aux points $x_1 \pm h, x_2 \pm h, \ldots, x_n \pm h$. Parzen (1962) a

proposé une généralisation de l'idée de Rosenblatt permettant, entre autres, de lisser davantage l'estimation. A la fonction K ci-dessus on substitue une fonction que l'on pourra choisir continue ou dérivable partout, propriété qui se transfère à la fonction $\widehat{f_n}$. En d'autres termes on fera entrer ou sortir les points x_i «en douceur» quand on déplace la fenêtre. Toutefois la fonction K est soumise aux conditions suivantes :

- K est positive (ou nulle)
- K est paire
- $\int_{\mathbb{R}} K(u)du = 1$.

Une telle fonction est alors appelée *noyau*. La première condition garantit que le poids $K\left(\frac{x-x_i}{h}\right)$ de chaque observation x_i reste positif ou nul, la deuxième que ce poids soit identique de part et d'autre de x. La troisième condition est une normalisation des poids de façon que $\widehat{f_n}$ soit bien une densité. En effet, avec le changement de variable $u = \frac{x-x_i}{h}$ on obtient :

$$\int_{\mathbb{R}} \widehat{f_n}(x)dx = \frac{1}{nh}\sum_{i=1}^{n}\int_{\mathbb{R}} K\left(\frac{x-x_i}{h}\right)dx = \frac{h}{nh}\sum_{i=1}^{n}\int_{\mathbb{R}} K\left(u\right)du = \frac{1}{n}\sum_{i=1}^{n} 1 = 1.$$

Notons qu'une fonction noyau est, en fait, une fonction de densité symétrique autour de zéro, donc de moyenne nulle (si elle existe).

Comme :

$$\frac{1}{nh}\int_{\mathbb{R}} K\left(\frac{x-x_i}{h}\right)dx = \frac{1}{n}\int_{\mathbb{R}} K(u)du = \frac{1}{n}\,,$$

on peut donner une interprétation concrète de $\widehat{f_n}$. Supposons, pour fixer les idées, que K ait pour support $[-1,+1]$. Alors $\widehat{f_n}$ est obtenue en remplaçant chaque observation x_i par une même petite «densité» (son aire étant réduite à $\frac{1}{n}$) de support $[x_i-h, x_i+h]$, puis en sommant ces petites densités. Cette vision correspond à un principe général de lissage de données discrètes qui consiste à faire «bouger» chaque donnée pour lui substituer un élément continu.

En pratique on impose comme condition supplémentaire que K décroisse de part et d'autre de zéro, dans l'idée naturelle de donner un poids plus faible aux observations au fur et à mesure qu'elles s'éloignent du centre de la fenêtre x. Ainsi les noyaux les plus usuels sont :

$K(u) = \frac{1}{2}$	si $u \in [-1,1]$	noyau de Rosenblatt
$K(u) = 1 - \lvert u\rvert$	si $u \in [-1,1]$	noyau triangulaire
$K(u) = \frac{3}{4}(1-u^2)$	si $u \in [-1,1]$	noyau d'Epanechnikov
$K(u) = \frac{15}{16}(1-u^2)^2$	si $u \in [-1,1]$	noyau de Tukey ou biweight
$K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$	$u \in \mathbb{R}$	noyau gaussien.

Les deux premiers ont l'avantage d'être simples, le noyau triangulaire étant continu partout et conduisant à une estimation $\widehat{f}_n$ continue. Le troisième doit sa notoriété à une propriété d'optimalité théorique mais sans grand intérêt pratique (voir plus loin le paragraphe «choix pratiques»). Le quatrième est, à notre sens, le plus intéressant car donnant une estimation dérivable partout, tout en étant simple à mettre en oeuvre. En fait il s'agit du noyau le plus simple parmi les noyaux de forme polynomiale dérivables partout. Ainsi il assure le lissage local de la fonction $\widehat{f}_n$. Ce noyau est d'une forme très proche du noyau gaussien et est donc préférable, ce dernier ayant un coût de calcul plus élevé du fait de son support infini (la «largeur de fenêtre» h devenant conventionnellement l'écart-type de la loi de Gauss). Notons que plus la valeur de h est élevée plus on élargit la fenêtre, ce qui a un effet de lissage global de $\widehat{f}_n$ plus important. Ceci est à rapprocher du choix de la largeur des intervalles pour l'histogramme.

Note 8.5 Le choix du type de noyau étant fixé, seul reste à effectuer le choix de h que nous envisagerons plus loin. Le problème du positionnement de la grille de l'histogramme (choix de a_0) n'existe pas ici. Pour ce dernier on peut aussi s'affranchir de ce choix en prenant la moyenne, en continu, des estimations obtenues par glissement continu de la grille de a_0 quelconque à $a_0 + h$. On obtient alors l'estimateur à noyau triangulaire (voir exercices).

Propriétés asymptotiques des estimateurs à noyaux Il existe peu de résultats à n fini et l'on doit se satisfaire de résultats asymptotiques. Reprenons l'expression générale d'un estimateur à noyau pour un échantillon aléatoire $X_1, X_2, \ldots, X_n$:

$$\widehat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Pour calculer le biais et la variance en un point x fixé posons :

$$Z_i = \frac{1}{h} K\left(\frac{x - X_i}{h}\right).$$

Ainsi la variable aléatoire $\widehat{f}_n(x)$ est la moyenne des Z_i qui, en tant que fonctions respectives des X_i, sont des variables aléatoires i.i.d.. Soit $Z = \frac{1}{h} K\left(\frac{x-X}{h}\right)$ la v.a. symbolisant la loi commune aux Z_i comme X symbolise la loi mère des X_i de densité f.

Calcul du biais

On a :

$$\begin{aligned} E(\widehat{f}_n(x)) &= E(\frac{1}{n}\sum_{i=1}^{n} Z_i) = E(Z) = \frac{1}{h}E\left(K\left(\frac{x-X}{h}\right)\right) \\ &= \frac{1}{h}\int_{\mathbb{R}} K\left(\frac{x-t}{h}\right) f(t)dt \\ &= \int_{\mathbb{R}} K(u)f(x+uh)du \quad \text{en posant } u = \frac{x-t}{h} \ . \end{aligned}$$

Comme $\int_{\mathbb{R}} K(u)du = 1$ le biais peut s'écrire :

$$E(\widehat{f}_n(x)) - f(x) = \int_{\mathbb{R}} K(u)[f(x+uh) - f(x)]du \ .$$

On voit que le biais résulte de l'écart de la valeur de la densité dans la fenêtre centrée sur x par rapport à sa valeur en x même. Si f était constante dans la fenêtre le biais serait nul, et de même si f était parfaitement linéaire en raison de la parité du noyau K. Comme pour l'histogramme le biais ne dépend pas de la taille de l'échantillon et ne peut être réduit à zéro qu'en faisant tendre h vers zéro. Prenons un développement de Taylor de f au voisinage de x :

$$f(x+uh) = f(x) + uhf'(x) + \frac{u^2h^2}{2}f''(x) + o(h^2) \ .$$

Le biais s'écrit :

$$\begin{aligned} E(\widehat{f}_n(x)) - f(x) &= hf'(x)\int_{\mathbb{R}} uK(u)du + \frac{h^2}{2}f''(x)\int_{\mathbb{R}} u^2K(u)du + o(h^2) \\ &= \frac{h^2}{2}f''(x)\int_{\mathbb{R}} u^2K(u)du + o(h^2) \end{aligned}$$

puisque K est paire. Pour h petit le biais dépend donc de $f''(x)$ et du moment d'ordre 2 du noyau. Le biais est du signe de $f"(x)$: si f est concave en x le biais est négatif, si elle est convexe le biais est positif. En particulier si x est un point où f est à un maximum le biais sera négatif. On sous-estime donc (en moyenne) la hauteur du maximum, ce que l'on peut comprendre intuitivement : la densité au voisinage de x étant plus faible il y a nécessairement un déficit de points dans la fenêtre. A l'inverse les minima éventuels seront surestimés. Par conséquent la méthode tend à écrêter les creux et les pics de la densité ce qui est un inconvénient majeur.

Calcul de la variance

On a :

$$V(\widehat{f}_n(x) = V(\frac{1}{n}\sum_{i=1}^{n} Z_i) = \frac{1}{n}V(Z) = \frac{1}{n}\left\{E(Z^2) - [E(Z)]^2\right\}$$

avec :

$$\frac{1}{n}E(Z^2) = \frac{1}{nh^2}\int_{\mathbb{R}}\left[K\left(\frac{x-t}{h}\right)\right]^2 f(t)dt$$
$$= \frac{1}{nh}\int_{\mathbb{R}}\left[K(u)\right]^2 f(x+uh)du \quad \text{(en posant } u = \frac{x-t}{h}\text{)}$$

et :

$$\frac{1}{n}[E(Z)]^2 = \frac{1}{n}\left[\int_{\mathbb{R}} K(u)f(x+uh)du\right]^2 .$$

Alors que le terme $\frac{1}{n}[E(Z)]^2$ tend bien vers zéro quand $n \to \infty$ on voit que le terme $\frac{1}{n}E(Z^2)$ ne tend vers zéro que si $nh \to \infty$. Par conséquent, pour que $\widehat{f}_n(x)$ converge vers $f(x)$ en moyenne quadratique les mêmes conditions sont nécessaires que pour l'histogramme : $n \to \infty, h \to 0, nh \to \infty$.

Le terme $\frac{1}{n}[E(Z)]^2$ est d'ordre $\frac{1}{n}$, ce que l'on note $O(\frac{1}{n})$. En utilisant le développement de Taylor : $f(x+uh) = f(x) + uhf'(x) + o(h)$, on obtient :

$$\frac{1}{n}E(Z^2) = \frac{1}{nh}f(x)\int_{\mathbb{R}}\left[K(u)\right]^2 du + O(\frac{1}{n}),$$

d'où :

$$V(\widehat{f}_n(x) = \frac{1}{nh}f(x)\int_{\mathbb{R}}\left[K(u)\right]^2 du + O(\frac{1}{n}).$$

Finalement l'erreur quadratique moyenne en x fixé est :

$$eqm(\widehat{f}_n(x)) = \frac{h^4}{4}[f''(x)]^2\left[\int_{\mathbb{R}} u^2K(u)du\right]^2 + \frac{f(x)}{nh}\int_{\mathbb{R}}\left[K(u)\right]^2 du$$
$$+ o(h^4) + O(\frac{1}{n}) .$$

Faisant abstraction des termes $o(h^4) + O(\frac{1}{n})$ négligeables dans les conditions de convergence, on voit que plus la largeur de fenêtre h est faible plus le biais diminue mais plus la variance augmente et, inversement, l'élargissement de la fenêtre augmente le biais et diminue la variance. Il existe un optimum (mais valable uniquement pour le point x) qui, comme pour l'histogramme, est obtenu en dérivant par rapport à h, soit :

$$h_{opt} = \left[\frac{f(x)\int_{\mathbb{R}}\left[K(u)\right]^2 du}{[f''(x)]^2\left[\int_{\mathbb{R}} u^2K(u)du\right]^2}\right]^{1/5} n^{-1/5}$$

et, en substituant h_{opt} dans la formule de l'expression asymptotique de l'e.q.m., celle-ci prend la forme $k(x)\,\nu(K)\,n^{-4/5}$ où $\nu(K)$ est une expression qui ne dépend que du choix du noyau et $k(x)$ est fonction de $f(x)$ et de $f''(x)$. Ainsi

la convergence est plus rapide que pour l'histogramme, étant d'ordre $n^{-4/5}$ au lieu de $n^{-2/3}$.

Jusqu'à présent nous avons raisonné à x fixé. Il est clair que ce qui nous intéresse est de connaître le comportement de l'estimateur $\widehat{f}_n$ de la fonction f globalement sur tout $\mathbb{R}$. Pour cela on considère, pour une réalisation donnée, son écart à f intégré sur tout $\mathbb{R}$ ce qui conduit, en prenant l'espérance mathématique de cet écart intégré, au critère d'*erreur quadratique intégrée moyenne* (e.q.i.m. ou MISE en anglais : *mean integrated square error*) :

$$eqim(\widehat{f}_n) = E\left(\int_{\mathbb{R}} \left[\widehat{f}_n(x) - f(x)\right]^2 dx\right).$$

Celle-ci se calcule aisément à partir des résultats précédents car, étant donné les conditions de régularité imposées à f et à K, il est licite d'intervertir les intégrations (l'une explicite, l'autre implicite dans le calcul de l'espérance mathématique) ce qui conduit à intégrer l'expression de l'e.q.m. en x fixé :

$$eqim(\widehat{f}_n) = \int_{\mathbb{R}} E\left(\left[\widehat{f}_n(x) - f(x)\right]^2\right) dx = \int_{\mathbb{R}} eqm(\widehat{f}_n(x))dx.$$

D'où :

$$eqim(\widehat{f}_n) = \frac{h^4}{4}\int_{\mathbb{R}} [f''(x)]^2 dx \left[\int_{\mathbb{R}} u^2 K(u)du\right]^2 + \frac{1}{nh}\int_{\mathbb{R}} [K(u)]^2\, du + o(h^4) + O(\frac{1}{n}).$$

Comme précédemment on trouve un h optimal qui est en $n^{-1/5}$ et une e.q.i.m. de la forme $g(f'')\,\nu(K)\,n^{-4/5}$. Le même critère aurait pu être appliqué à l'histogramme, la vitesse de convergence étant également conservée en $n^{-2/3}$.

Ainsi, en tant qu'estimateur fonctionnel un estimateur à noyau converge plus vite vers la vraie densité f que l'histogramme. Mais ce résultat repose sur un choix optimal très théorique (puisque dépendant de l'inconnue f'') et de conditions de convergence artificielles. C'est pourquoi nous considérons maintenant les aspects pratiques.

Choix pratiques Le praticien doit effectuer deux choix : celui du noyau K et celui de la fenêtre h.

Le choix de K s'avère être relativement indifférent pour ce qui concerne le critère de l'e.q.i.m. Ceci a pu être constaté par calcul direct ou par simulation sur une grande variété de lois mères et est d'ailleurs confirmé sur l'expression asymptotique ci-dessus. En effet la valeur minimale de $\nu(K)$ est atteinte avec le noyau d'Epanechnikov. Cette valeur est 0,349 alors qu'elle est égale à 0,351 pour le biweight et 0,369 pour le noyau de Rosenblatt (voir exercices). Par conséquent

le biweight (noyau de Tukey) doit être recommandé pour son avantage de lissage local évoqué plus haut.

Reste le choix difficile de h pour lequel diverses méthodes ont été proposées, aucune ne donnant satisfaction de façon universelle. Sans aborder dans le détail ce vaste sujet mentionnons trois approches.

Deheuvels (1977) a suggéré de prendre la valeur optimale vue ci-dessus : $m(f'')\,\nu(K)\,n^{-4/5}$, en calculant $m(f'')$ sur f gaussienne. Toutefois on n'échappe pas à l'estimation de la variance σ de la loi de Gauss que l'on effectue naturellement par la variance empirique.

Une deuxième approche dérivée d'une procédure de type rééchantillonnage dite de *validation croisée* a été étudiée initialement par Marron (1987) et est souvent adoptée par les logiciels, car de portée plus générale. Elle consiste à choisir la valeur de h qui maximise l'expression :

$$\prod_{i=1}^{n} \widehat{f}_{n,h}^{-i}(x_i)$$

où $\widehat{f}_{n,h}^{-i}$ est l'estimation de densité effectuée avec une valeur h en omettant la i-ème observation. On maximise ainsi globalement les densités attribuées aux observations x_i à la manière du maximum de vraisemblance. Pour que l'évaluation en x_i ne soit pas influencée par la valeur de x_i elle-même, on élimine celle-ci du calcul.

Pour notre part nous proposons de choisir, avec le noyau biweight :

$$h = 0{,}75 \min_i [x_{(i+[\frac{n}{2}])} - x_{(i)}] \left(\frac{n}{100}\right)^{-1/5}.$$

Ceci résulte du fait que, pour une diversité de lois et pour $n = 100$, la valeur optimale reste proche de 0,75 fois la largeur de l'intervalle de probabilité 0,5 autour du mode. Cette méthode est simple à mettre en oeuvre et effectue généralement un lissage adéquat (voir Lejeune, 1982).

La plupart des propositions de choix de h reposent sur l'optimisation de l'e.q.i.m. (ou sur l'un des critères mentionnés plus loin, mais aucun n'est la panacée) et, par expérience, on constate souvent qu'elles ne fournissent pas nécessairement une estimation graphiquement satisfaisante, laissant subsister des variations locales (tendance à sous-lisser). La méthode la plus sûre reste donc celle des essais et erreurs où, partant d'une valeur de toute évidence trop faible de h donnant des fluctuations locales indésirables, on augmente progressivement cette valeur jusqu'au seuil de disparition de telles fluctuations.

Par calcul direct ou par simulation on constate que, même pour d'assez grands échantillons, la valeur de h effectivement optimale (au sens de l'e.q.i.m. mais aussi d'autres critères d'erreur) reste étonnamment élevée. Ainsi en prenant, simplement à titre indicatif, une loi mère $\mathcal{N}(0\,;1)$ la valeur de h optimale

est de 1,11 pour $n = 100$ et vaut encore 0,70 pour $n = 1\,000$. Ceci restreint fortement la validité des expressions asymptotiques établies avec h tendant vers zéro. De plus avec un tel choix il faut s'attendre à un écrêtement non négligeable des extrema de la densité.

Même si les expressions asymptotiques sont à prendre avec précaution, elles permettent de vérifier sur diverses lois que la méthode des noyaux, outre le fait qu'elle peut donner une estimation lisse, est nettement plus efficace, au sens de l'e.q.i.m., que l'histogramme. En se plaçant, par exemple, aux valeurs respectives optimales de h avec $n = 100$, pour une loi mère $\mathcal{N}(0\,;1)$, l'e.q.i.m. asymptotique de l'histogramme est 2,5 fois plus élevée que celle obtenue par noyau biweight. Lorsque n s'accroît ce rapport augmente, ce qui correspond aux différences de vitesses de convergence : pour $n = 1\,000$ il vaut 3,4 (voir exercices).

Remarques diverses On a vu que le biais était de nature à écrêter les extrema de f ce qui est particulièrement fâcheux s'agissant de points caractéristiques de la densité (dont son ou ses modes). On peut réduire le biais, et notamment ce phénomène, en relâchant la contrainte de positivité du noyau.

En effet en choisissant un noyau K tel que $\int_{\mathbb{R}} u^2 K(u)du = 0$ on élimine, dans l'expression asymptotique du biais le terme en h^2 : $\frac{h^2}{2} f''(x) \int_{\mathbb{R}} u^2 K(u)du$. En poursuivant le développement de Taylor jusqu'à l'ordre 4 l'expression asymptotique du biais devient $\frac{h^4}{4!} f^{(4)}(x) \int_{\mathbb{R}} u^4 K(u)du + o(h^4)$ où $f^{(4)}$ est la dérivée à l'ordre 4 de f. Ainsi, si f reste proche d'un polynôme du deuxième ou troisième degré au voisinage de x, le biais sera pratiquement réduit à zéro, ce qui permet de mieux coller aux extrema. Un noyau vérifiant $\int_{\mathbb{R}} u^2 K(u)du = 0$ (et par la parité, forcément $\int_{\mathbb{R}} u^3 K(u)du = 0$) est appelé *noyau d'ordre 4*. Le noyau d'ordre 4 dérivable partout de type polynomial le plus simple est :

$$K(u) = \frac{105}{64}(1-u^2)^2\ (1-3u^2) \quad \text{si } u \in [-1,1] \quad (0 \text{ sinon}),$$

qui est une modification du biweight et dont la figure 8.1 montre qu'il a des plages négatives sur les extrémités. Ce faisant la variance est accrue par rapport au biweight, mais cet accroissement est largement compensé par la diminution du biais pour le critère d'e.q.i.m. (à titre d'exemple, pour une loi $\mathcal{N}(0\,;1)$ le gain global est de l'ordre de 25% avec des largeurs de fenêtre autour de l'optimum).

Toutefois le prix à payer pour réduire le biais est le fait que $\widehat{f}_n$ n'est plus nécessairement positive ou nulle. Étant donné le faible poids des plages négatives dans le noyau (voir figure 8.1) il s'agira d'effets de bord : les zones négatives de $\widehat{f}_n$ seront limitées aux extrémités, où les observations se font rares, et seront de très faible ampleur. Néanmoins on sera contraint de « rectifier » $\widehat{f}_n$ sur ces bords. En raison de cet inconvénient, même marginal, le noyau d'ordre 4 n'a pas le succès qu'il mériterait pourtant. La figure 8.2 est un exemple obtenu avec le noyau proposé ci-dessus.

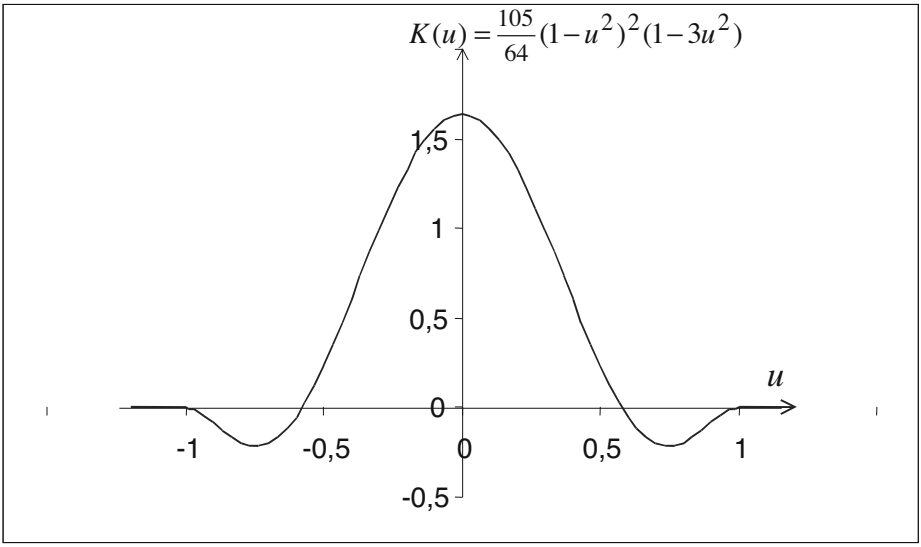


Figure 8.1 - Noyau d'ordre 4 dérivé du biweight

Comme nous l'avons déjà indiqué d'autres critères d'erreur que l'erreur quadratique intégrée ont été étudiés, notamment :

$$\int_{\mathbb{R}} \left| \widehat{f}_n(x) - f(x) \right| dx$$

$$\sup_x \left| \widehat{f}_n(x) - f(x) \right|$$

dont on a montré la convergence en probabilité vers zéro avec le même type de conditions que précédemment (citons à ce propos les travaux de Devroye et Györfi, 1985).

D'autres approches que la méthode des noyaux ont été proposées : séries orthogonales, splines, maximum de vraisemblance pénalisé, plus proches voisins (*nearest neighbour*), ondelettes, etc. Globalement on peut dire qu'elles ne donnent pas des résultats significativement meilleurs que la méthode des noyaux. Aucune ne peut se soustraire à l'incontournable problème du choix d'un paramètre de lissage, explicite ou non.

8.5.3 L'estimation de la fonction de répartition

La démarche que nous allons suivre est calquée sur celle de la densité. Partant de la solution classique de la statistique descriptive, à savoir la fonction de répartition empirique, nous étudierons les possibilités de lissage en vue d'une amélioration.

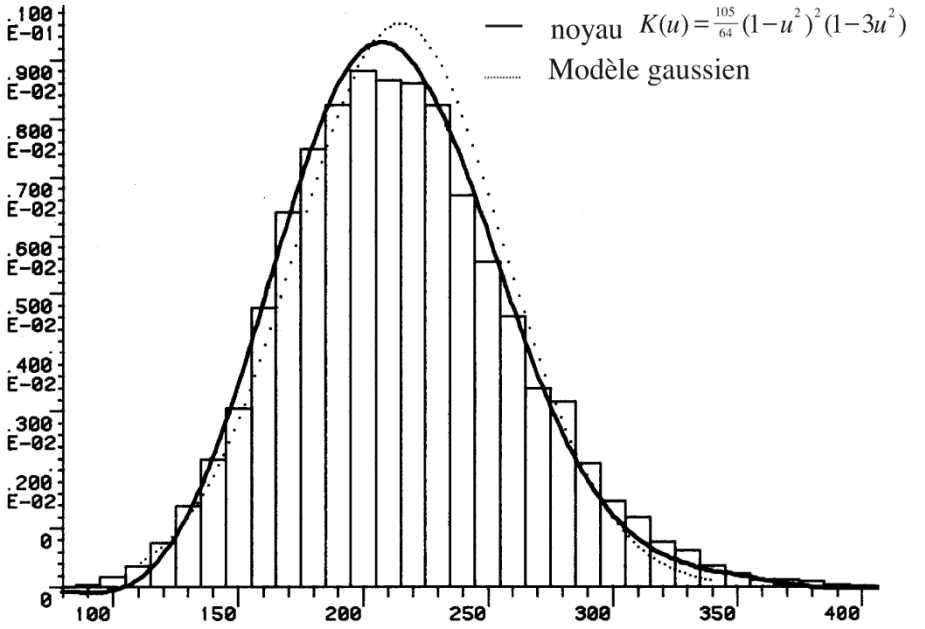


Figure 8.2 - Estimation par noyau : données de taux de cholestérol de 3 200 hommes (*source : projet FNRS-Suisse sur la prévention des maladies cardio-vasculaires)*

La fonction de répartition empirique

Rappelons sa définition donnée en section 5.7 :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i) \text{ pour tout } x \in \mathbb{R},$$

où $I_{(-\infty,x]}$ est la fonction indicatrice de l'intervalle $(-\infty, x]$. Pour une réalisation $x_1, x_2, \ldots, x_n$, c'est une fonction en escalier s'élevant de $1/n$ à chaque rencontre d'une valeur x_i. Pour x fixé on a vu que la statistique $nF_n(x)$ suit une loi binomiale $\mathcal{B}(n, F(x))$. Pour n grand, par l'approximation gaussienne d'une binomiale (voir section 5.8.3), $nF_n(x)$ suit approximativement une loi $\mathcal{N}(nF(x), nF(x)(1-F(x)))$ et on a donc :

$$F_n(x) \underset{approx}{\rightsquigarrow} \mathcal{N}\left(F(x)\,,\, \frac{F(x)(1-F(x))}{n}\right).$$

Les deux paramètres de cette loi de Gauss correspondent à la moyenne et la variance de $F_n(x)$. Pour x fixé, $F_n(x)$ est donc un estimateur sans biais et convergent de $F(x)$. Notons que $F_n(x)$ est la moyenne d'une suite de variables

aléatoires i.i.d. $I_{(-\infty,x]}(X_i)$ (elles sont indépendantes comme fonctions respectives des X_i) toutes issues de la loi de Bernoulli de paramètre $p = F(x)$. Ainsi la loi des grands nombres (théorème 5.3) s'applique et $F_n(x)$ converge aussi presque sûrement vers $F(x)$. L'approximation gaussienne ci-dessus permet de construire un intervalle de confiance approché sur le modèle de celui de la section 7.4.5 pour le paramètre p d'une loi de Bernoulli.

Dans cette section nous nous intéressons aux propriétés de F_n en tant qu'estimateur fonctionnel de F. La première proposition qui suit a déjà été mentionnée plus haut dans le cadre du rééchantillonnage (section 8.4).

Proposition 8.5 *La fonction de répartition F_n est l'estimateur fonctionnel du maximum de vraisemblance pour F.*

Cette proposition démontrée par Kiefer et Wolfowitz (1956) mérite d'être commentée dans la mesure où, jusqu'à présent, l'estimateur du maximum de vraisemblance (EMV) n'a été défini que dans le cadre paramétrique (définition 6.11). Le principe reste le même : il s'agit de donner aux valeurs observées $x_1, x_2, \cdots, x_n$ la plus forte densité ou fonction de probabilité. Dans le cas continu on doit rechercher pour quelle fonction F l'expression de la fonction de vraisemblance de $F : L(F) = \prod_{i=1}^{n} F'(x_i)$, est maximisée. Pour une loi discrète $F'(x_i)$ doit être remplacé par $F(x_i) - F(x_i^-)$ correspondant à la fonction de probabilité en x_i (voir section 1.3). En fait on a avantage à rester dans la plus grande généralité, n'ayant aucun a priori sur la nature de la loi et considérant une maximisation sur l'ensemble $\mathcal{F}$ des fonctions de répartition englobant le cas discret, continu ou mixte. $\mathcal{F}$ est donc l'ensemble des fonctions répondant aux conditions nécessaires et suffisantes d'une fonction de répartition (croissance sur $\mathbb{R}$ de 0 à 1, continuité à droite en chaque point). Cette approche générale nécessite des connaissances au-delà du niveau de cet ouvrage et nous admettrons donc que la solution du problème de maximisation sur $\mathcal{F}$ est F_n. L'intérêt d'en rester à une approche générale tient au fait qu'une solution simple existe (si l'on met une contrainte de continuité pour F le problème devient difficile) et qu'elle est naturelle dans la mesure où il s'ensuit que l'EMV de toute caractéristique de la loi mère s'exprimant comme une espérance mathématique d'une fonction $g(X)$ devient alors $\frac{1}{n}\sum_{i=1}^{n} g(Xi)$, le cas le plus simple étant celui de la moyenne μ dont l'EMV est $\overline{X}$ (voir note 8.2).

Les propriétés de $F_n(x)$ pour x fixé ont été établies, notamment la convergence vers $F(x)$. Le théorème suivant (que nous admettrons) est essentiel car il montre la convergence **uniforme**, pour tout $x \in \mathbb{R}$, de F_n vers F.

Théorème 8.1 *(Glivenko-Cantelli) Soit un échantillon aléatoire $X_1, X_2, \ldots, X_n$ issu de la loi de fonction de répartition F et F_n sa fonction de répartition empirique. Alors, quand $n \longrightarrow \infty$:*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p.s.} 0\,.$$

Pour voir les choses concrètement, ce théorème nous dit que l'on peut être assuré que l'écart maximal entre F_n et F va tendre vers 0 si l'on augmente la taille de l'échantillon à l'infini ou encore que partout, simultanément, la fonction de répartition empirique va se rapprocher de la vraie fonction de répartition. De plus, le théorème suivant donne le comportement asymptotique de l'écart maximal entre F_n et F.

Théorème 8.2 *(Kolmogorov-Smirnov) Soit la variable aléatoire :*

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

Alors, pour $x > 0$, *on a :*

$$P(\sqrt{n} D_n < x) \underset{n \to \infty}{\longrightarrow} 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2(kx)^2}.$$

En d'autres termes $\sqrt{n} D_n$ tend en loi vers une v.a. à valeurs positives (car D_n est nécessairement positive) de fonction de répartition $G(x)$ égale à l'expression limite ci-dessus, laquelle ne dépend pas de F. En fait même pour n fini **la loi de** D_n **ne dépend pas de** F et, de ce fait, elle a été tabulée. A partir de $n = 40$ l'approximation par $G(x)$ est correcte à 10^{-2} près.

Ce résultat permet de donner une *bande de confiance* approchée pour F. En effet, soit $g_{0,95}$ le quantile d'ordre 0,95 de $G(x)$, on a :

$$P(\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| < g_{0,95}) \simeq 0,95\,.$$

Mais l'événement $(\sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| < g_{0,95})$ équivaut à l'événement

$$(\sqrt{n}[F_n(x) - F(x)| < g_{0,95}]\,, \text{ pour tout } x)$$

ou :

$$(F_n(x) - \frac{g_{0,95}}{\sqrt{n}} < F(x) < F_n(x) + \frac{g_{0,95}}{\sqrt{n}}, \text{ pour tout } x).$$

On a donc une procédure qui garantit, a priori avec une probabilité 0,95, que $F(x)$, pour tout x, soit compris dans l'intervalle :

$$F_n(x) \pm \frac{g_{0,95}}{\sqrt{n}}.$$

Pour une réalisation, la bande autour de F_n ainsi dessinée sera une région de confiance à 95% approchée pour F. On peut même établir une bande exacte en lisant dans une table le quantile 0,95 de la loi exacte de D_n. A partir de $n = 40$ on peut utiliser l'expression asymptotique qui peut se réduire, disons si $x >$0,8, pratiquement à $1 - 2e^{-2x^2}$ en ne gardant que le premier terme de

la somme. Ainsi $g_{0,95}$ est approximativement défini par la valeur de x telle que $1-2e^{-2x^2} = 0{,}95$ soit $x = 1{,}36$. Pour n assez grand la bande de confiance à 95 % est donc $F_n(x) \pm \frac{1{,}36}{\sqrt{n}}$.

Le théorème 8.2 trouvera plus loin une application très répandue pour tester un modèle de loi (voir le test de Kolmogorov-Smirnov en section 10.4.1).

Lissage de F_n

Nous envisageons maintenant le lissage de F_n pour le cas où l'on sait pouvoir se restreindre à une fonction de répartition dérivable jusqu'à un certain ordre. Il existe, comme pour la densité, une série de solutions, mais nous ne présenterons que celle qui fait le pendant de l'estimateur à noyau de la densité.

Considérons l'estimation de F obtenue en intégrant l'estimation par noyau de la densité f :

$$\begin{aligned}\widehat{F}_n(x) &= \int_{-\infty}^{x} \widehat{f}_n(t)dt = \int_{-\infty}^{x} \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{t-x_i}{h}\right) dt \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{x} K\left(\frac{t-x_i}{h}\right) dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-x_i}{h}} K(v)\, dv \quad \text{en posant } v = \frac{t-x_i}{h}\,.\end{aligned}$$

Définissons le *noyau intégré* :

$$H(u) = \int_{-\infty}^{u} K(u)\, du\,,$$

alors on a :

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x-x_i}{h}\right)\,.$$

Comme $K(u)$ était de la forme d'une fonction de densité, $H(u)$ est de la forme d'une fonction de répartition.

Pour voir l'analogie avec la fonction de répartition empirique F_n, rappelons que $F_n(x)$ est la moyenne des indicatrices $I_{(-\infty,x]}(x_i)$. Or :

$$I_{(-\infty,x]}(x_i) = I_{(-\infty,0]}(x_i - x) = I_{(0,+\infty,]}(\frac{x-x_i}{h}) = \left\{ \begin{array}{l} 0 \text{ si } x < x_i \\ 1 \text{ si } x \geq x_i \end{array} \right. .$$

F_n est donc de la forme de $\widehat{F}_n(x)$, avec une fonction $H(u)$ très particulière, donnant le saut brutal de 0 à 1 en $u = 0$. L'apport d'une fonction plus souple répond au même principe général de lissage que celui évoqué pour la densité,

à savoir qu'on effectue un passage de 0 à 1 en douceur, étalé entre $x_i - h$ et $x_i + h$ autour de x. On remarquera que du fait que H est une primitive de K elle est continue et $\widehat{F}_n$ est donc également continue.

Par intégration du noyau de Rosenblatt, le plus simple, et de celui de Tukey préconisé pour la densité, on obtient :

$$H_1(u) = \begin{cases} 0 & \text{si } u \leq -1 \\ \frac{1}{2}(u+1) & \text{si } -1 < u < +1 \\ 1 & \text{si } +1 \leq u \end{cases} \qquad \text{(Rosenblatt intégré)}$$

$$H_2(u) = \begin{cases} 0 & \text{si } u \leq -1 \\ \frac{1}{16}(8 + 15u - 10u^3 + 3u^5) & \text{si } -1 < u < +1 \\ 1 & \text{si } +1 \leq u \end{cases} \quad \text{(Tukey intégré)}.$$

Lorsqu'on examine le graphe obtenu avec différents noyaux on constate que la différence est imperceptible. Ceci s'explique par le fait que l'estimation d'une fonction de répartition est fortement contrainte par la condition de croissance de 0 à 1, et par sa continuité. De ce fait le problème est beaucoup plus simple que pour la densité. En particulier la croissance implique de faibles courbures et donc peu ou pas de problème de biais, contrairement à la densité. Il n'y a donc pas d'avantage tangible à utiliser des noyaux ou autres instruments de lissage sophistiqués et nous préconisons donc l'emploi du noyau H_2. Notons bien, toutefois, que les estimations de densités obtenues par dérivation seront, quant à elles, très sensibles aux variations jugées mineures pour la fonction de répartition. Malgré ce constat il n'est pas inutile d'examiner le biais et la variance, de façon asymptotique, comme cela a été fait pour la densité.

Biais et variance On démontre les résultats suivants (voir Lejeune et Sarda, 1992) par des développements similaires à ceux de la densité. Pour un noyau K symétrique de support [-1,+1] on a, en x fixé :

$$E(\widehat{F}_n(x)) - F(x) = \frac{h^2}{2} f'(x) \int_{\mathbb{R}} u^2 K(u) du + o(h^2)\,,$$

$$V(\widehat{F}_n(x)) = \frac{1}{n}\left\{ F(x)[1 - F(x)] + h f(x) \left[\int_{-1}^{+1} H^2(u) du - 1 \right] + o(h) \right\}.$$

Alors que la fonction de répartition empirique F_n est sans biais, le lissage ne peut éviter d'introduire un certain biais. Les simulations à n fini montrent toutefois que ce biais reste très faible. En particulier on voit sur l'expression asymptotique qu'à $o(h^2)$ près il s'annule aux extrema de la densité (donc aux modes) qui correspondent à des points d'inflexion pour F. Pour la variance on retrouve dans le premier terme de son expression asymptotique la variance de F_n. Par conséquent on gagne sur la variance de F_n si le deuxième terme est négatif, soit $\int_{-1}^{+1} H^2(u) du < 1$, ce qui est vérifié pour les noyaux intégrés

courants. Dans le cas du noyau de Rosenblatt $\int_{-1}^{+1} H^2(u)du = \frac{2}{3}$, la variance décroît donc de $\frac{h}{3n}f(x)$ et le biais vaut $\frac{h^2}{6}f'(x)$ (à $o(h^2)$ près). Pour ce qui concerne l'erreur quadratique moyenne elle sera améliorée si la diminution de variance compense le biais au carré. Pour estimer, par exemple, le mode d'une loi qui serait une loi de Gauss, il n'y a pas de biais en raison de la symétrie en ce point et la variance diminue de 15 %.

Ici comme pour la densité se pose le problème de la largeur de fenêtre optimale. Il est toutefois moins crucial en raison de la plus faible sensibilité de l'estimation à ce paramètre de lissage.

Note 8.6 On peut penser que l'estimation d'une caractéristique $\omega(F)$ par $\omega(\widehat{F}_n)$ puisse être meilleure que la simple version empirique $\omega(F_n)$. Encore faut-il choisir une valeur de h appropriée (la valeur optimale pour ce problème étant alors généralement plus faible qu'avec l'objectif de minimisation de l'erreur quadratique intégrée moyenne). Par ailleurs on a avantage à utiliser une version «rétrécie» de $\widehat{F}_n$ qui conserve la variance empirique s^2. Elle s'obtient en remplaçant x par $\sqrt{1+\frac{h^2}{s^2}}\,x$ dans l'expression de $\widehat{F}_n(x)$. Silverman (1987) montre que pour la plupart des lois, le moment simple d'ordre 6 est mieux estimé par une version lisse, alors que l'amélioration n'est pas systématique pour les moments d'ordres inférieurs.

Pour approfondir l'estimation fonctionnelle on pourra consulter l'ouvrage méthodologique très complet de Simonoff (1996) ou, pour les aspects mathématiques, celui de Bosq et Lecoutre (1987).

8.6 Exercices

Exercice 8.1 Générer 200 observations de loi lognormale $\mathcal{LN}(0\,;1)$ (aide : dans un tableur du type EXCEL générer 200 observations de loi $\mathcal{N}(0\,;1)$ et les transformer par e^x).

Donner une estimation ponctuelle et par intervalle pour la médiane de la loi. L'intervalle contient-il la vraie valeur ? Recommencer pour estimer le quantile d'ordre 0,90.

Exercice 8.2 (Adapté de Mosteller et Tukey, 1977) Soit les valeurs 0,1 0,1 0,1 0,1 0,4 0,5 1,0 1,1 1,3 1,9 1,9 4,7. Donner une estimation du jackknife de l'écart-type de la loi mère ayant généré ces observations, fondée sur la statistique S. Donner un intervalle de confiance pour cet écart-type.

Exercice 8.3 Montrer que les pseudo-valeurs du jackknife fondées sur la variance empirique $\widetilde{S}_n^2$ sont $\frac{n}{n-1}(X_i - \overline{X})^2$, $i = 1, \ldots, n$.

Exercice 8.4 Vérifier que l'estimateur du jackknife appliqué à la moyenne empirique redonne la moyenne empirique. Quelles sont les pseudo-valeurs ?

Exercice 8.5 * Dans les notations de la section 8.5.2 concernant l'histogramme, écrire l'estimation $\widehat{f}_n(x)$ sous la forme d'une somme d'indicatrices. Dans le cas d'une grille régulière $\{a_k\}$ de largeur d'intervalle h, déterminer $\widehat{f}_n(x)$ pour un déplacement de la grille $\{a_k + t\}$ où $t > 0$ et $t < h$. Calculer la valeur moyenne de $\widehat{f}_n(x)$ quand t varie de 0 à h. Montrer qu'on obtient ainsi une estimation par noyau triangulaire (voir note 8.5).

Exercice 8.6 Dans un tableur du type EXCEL générer 50 observations de loi $\mathcal{N}(0\,;1)$. Estimer $f(0)$ par un noyau biweight avec $h = 1$. Comparer à la vraie valeur. Recommencer la procédure plusieurs fois pour confirmer le type de biais en présence.

Exercice 8.7 Dans un tableur du type EXCEL générer 50 observations de loi $\mathcal{N}(0\,;1)$. Estimer $f(0)$ par un noyau biweight avec $h = 0{,}5\,;\ 0{,}75\,;\ 1\,;\ 1{,}25\,;\ 1{,}5$ pour apprécier la variabilité des estimations ainsi obtenues.

Quelle est la valeur de h asymptotiquement optimale ?

Exercice 8.8 Dans l'expression asymptotique avec h optimal de l'erreur quadratique intégrée moyenne de l'estimateur à noyau de la densité, déterminer l'expression $\nu(K)$ qui ne dépend que du noyau. Calculer et comparer $\nu(K)$ pour le noyau d'Epanechnikov, le noyau de Rosenblatt et le noyau de Tukey.

Exercice 8.9 A partir de l'expression asymptotique de l'erreur quadratique intégrée moyenne de l'estimateur à noyau de la densité établir, dans le cas du noyau de Tukey et pour une loi mère $\mathcal{N}(\mu, \sigma^2)$, que la valeur de h optimale est $h_{opt} \simeq 2{,}78\,\sigma\, n^{-1/5}$ et que l'e.q.i.m. correspondante est environ $0{,}321\,\sigma^{-1} n^{-4/5}$ (aide : on utilisera l'expression de h_{opt} établie à l'exercice précédent et les valeurs numériques utiles concernant le noyau biweight. Par ailleurs on peut établir que $\int_{\mathbb{R}} [f''(x)]^2 dx = \frac{3\sigma^{-5}}{8\sqrt{\pi}}$ pour la loi de Gauss).

Montrer de même pour l'histogramme que la valeur optimale de h donnée par la formule asymptotique est d'environ $3{,}49\,\sigma\, n^{-1/3}$ avec une e.q.i.m. correspondante de $0{,}430\,\sigma^{-1} n^{-2/3}$ (aide : on prendra comme e.q.i.m. l'expression intégrée de l'e.q.m. de la proposition 8.4, soit $\frac{h^2}{12} \int_{\mathbb{R}} [f'(x)]^2 dx \ + \frac{1}{nh}$. On peut établir que $\int_{\mathbb{R}} [f'(x)]^2 dx = \frac{\sigma^{-3}}{4\sqrt{\pi}}$ pour la loi de Gauss).

Exercice 8.10 *Établir la formule du biais pour $\widehat{F}_n(x)$, estimateur à noyau intégré de F :

$$E(\widehat{F}_n(x)) - F(x) = \tfrac{h^2}{2} f'(x) \int_{\mathbb{R}} u^2 K(u) du + o(h^2)$$

Aide : utiliser une intégration par partie pour introduire K.

Chapitre 9

Tests d'hypothèses paramétriques

9.1 Introduction

Les tests statistiques constituent une approche décisionnelle de la statistique inférentielle. Un tel test a pour objet de décider sur la base d'un échantillon si une caractéristique de la loi mère (ou de la population) répond ou non à une certaine spécification que l'on appelle *hypothèse*, par exemple : la moyenne de la loi est supérieure à 10. Ces spécifications peuvent avoir diverses provenances : normes imposées, affirmations faites par un tiers (par exemple le fabricant d'un produit), valeurs cruciales de paramètres de modèles, etc.

Dans le cadre paramétrique où nous nous situerons initialement, les hypothèses portent sur le paramètre inconnu θ ou sur une fonction de ce paramètre $h(\theta)$ correspondant à une caractéristique d'intérêt de la loi. Dans le cas simple d'un espace paramétrique $\Theta \subseteq \mathbb{R}$, l'hypothèse spécifiera une valeur ou un intervalle de valeur pour θ (ou pour $h(\theta)$). Alors qu'un intervalle de confiance indique l'ensemble des valeurs plausibles, un test décidera si tel ensemble de **valeurs spécifiées** est plausible ou non. Bien que conceptuellement distinctes ces deux démarches reposent sur les mêmes bases mathématiques et de ce fait nous reprendrons de nombreux éléments du chapitre 7. Nous verrons d'ailleurs en fin de chapitre que, pour tout test, on peut établir une équivalence avec un intervalle de confiance. D'une façon générale il est plus facile de construire un test que de construire un intervalle de confiance. Aussi la multiplicité des tests justifiera-t-elle que nous approfondissions les problèmes d'optimalité. En outre on pourra utiliser la propriété d'équivalence pour définir des procédures d'intervalles de confiance dérivées de tests.

Les tests statistiques permettent d'aborder une grande variété d'hypothèses au-delà du test d'une hypothèse portant sur un paramètre. Par exemple : la

comparaison de plusieurs lois (ou populations), l'existence de liens entre plusieurs variables aléatoires, l'adéquation d'un modèle, etc. C'est au chapitre 10 que nous aborderons plus particulièrement des hypothèses de nature plus complexe, notamment dans le cadre non paramétrique.

Redéfinissons le cadre paramétrique. La loi observée est réputée appartenir à une famille de lois décrite par la famille de densités de probabilité (respectivement de fonctions de probabilité) $\{f(x;\theta);\theta \in \Theta\}$, la forme fonctionnelle f étant connue et seul le paramètre θ étant inconnu. Θ est l'espace paramétrique et il est inclus dans $\mathbb{R}^k$ où k est la dimension du paramètre θ. La fonction de répartition est notée $F(x;\theta)$, l'échantillon est $X_1, X_2, \cdots, X_n$ et X désignera la v.a. symbolisant la loi mère de l'échantillon.

Dans l'approche paramétrique la plus générale un test statistique consiste à décider d'accepter ou de rejeter une hypothèse spécifiant que θ appartient à un ensemble de valeurs Θ_0. Cette hypothèse de référence est appelée *hypothèse nulle* et est notée H_0. A contrario on définit *l'hypothèse alternative*, notée H_1, pour laquelle θ appartient à $\Theta_1 = \Theta - \Theta_0$ où $\Theta - \Theta_0$ dénote le complémentaire de Θ_0 par rapport à Θ. En bref on identifiera cette situation en écrivant que l'on teste :

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta_1,$$

le mot vs. étant l'abréviation du latin *versus*. Suivant la nature de Θ_0 et de Θ_1 on distinguera trois cas :

- hypothèse nulle simple et alternative simple où $\Theta = \{\theta_0, \theta_1\}$:
  $H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta = \theta_1$
- hypothèse nulle simple et alternative multiple :
  $H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0$
- hypothèse multiple et alternative multiple :
  $H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta_1.$

Pour une hypothèse nulle ou une hypothèse alternative *multiple* il est sous-entendu qu'il y a plusieurs valeurs possibles de θ. Nous commencerons par le premier cas qui, s'il est en réalité peu fréquent, permet de poser simplement les notions essentielles et d'établir des résultats qui pourront être étendus aux autres situations. Dans ce chapitre, comme pour les intervalles de confiance au chapitre 7, nous introduirons tout d'abord la théorie générale pour présenter ensuite les tests paramétriques classiques.

9.2 Test d'une hypothèse simple avec alternative simple

L'espace paramétrique Θ ne comprend donc que deux valeurs θ_0 et θ_1, la valeur θ_0 étant la valeur spécifiée à tester, i.e. :

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta = \theta_1.$$

Un test pour H_0 est **une règle de décision** fondée sur la valeur réalisée t d'une statistique T appelée *statistique de test.* Sauf exception la statistique T sera à valeurs dans $\mathbb{R}$, nous le supposerons implicitement. La règle est comme suit :

- si $t \in A$ (une partie de $\mathbb{R}$) on accepte H_0 ,
- si $t \in \overline{A}$ (partie complémentaire) on rejette H_0 .

La région A, qui est généralement un intervalle, sera appelée *région d'acceptation* et $\overline{A}$ *région de rejet.*

Une telle règle de décision recèle deux types d'erreur possibles du fait que la vraie valeur du paramètre est inconnue :

- rejeter H_0 alors qu'elle est vraie (i.e. $\theta = \theta_0$) : *erreur de première espèce,*
- accepter H_0 alors qu'elle est fausse (i.e. $\theta = \theta_1$) : *erreur de deuxième espèce.*

Étant donné que la décision se fonde sur un résultat d'origine aléatoire on caractérisera chaque erreur par sa probabilité. En théorie de la décision une probabilité d'erreur est appelée *risque,* d'où les définitions suivantes.

Définition 9.1 *On appelle* ***risque de première espèce*** *la valeur α telle que :*

$$\alpha = P_{\theta_0}(T \in \overline{A}),$$

c'est-à-dire la probabilité de rejeter H_0 alors qu'elle est vraie.

Il est usuel de noter cette probabilité $P(T \in \overline{A} \,|\, H_0)$ même s'il ne s'agit pas là d'une probabilité conditionnelle et, dorénavant, nous adopterons cette notation commode. Le risque de première espèce est aussi appelé en bref risque α.

Définition 9.2 *On appelle* ***risque de deuxième espèce*** *la valeur β telle que :*

$$\beta = P_{\theta_1}(T \in A),$$

c'est-à-dire la probabilité d'accepter H_0 alors que H_1 est vraie.

Ici également on notera cette probabilité $P(T \in A \,|\, H_1)$ et on parlera de risque β.

Les deux risques sont interdépendants puisque l'un repose sur A et l'autre sur son complémentaire $\overline{A}$. Par le choix de A ou de $\overline{A}$ on peut donc vouloir contrôler l'un ou l'autre, mais pas les deux. Dans un test statistique on privilégie en fait le risque α que l'on se fixe a priori et le plus souvent on prend $\alpha = 0{,}05$. C'est pourquoi la valeur α est aussi appelée le *niveau* ou *niveau de signification*[1] du test. Ce niveau ayant été choisi il s'agit de déterminer une région de rejet $\overline{A}$

[1] Ce terme de signification est à rapprocher de la notion de test de significativité présentée en fin de section 9.4.2.

telle que, «sous H_0», la probabilité que T «tombe» dans $\overline{A}$ soit effectivement égale à α. On voit ainsi que **la loi de la statistique de test doit être parfaitement connue sous** H_0. La construction d'un test consiste donc à rechercher une statistique pertinente (nous expliciterons plus loin ce que nous entendons par là) dont on connaît la loi sous H_0. La région de rejet étant ainsi déterminée, la région d'acceptation l'est aussi et donc également le risque de deuxième espèce β. Il est essentiel de garder à l'esprit que dans une procédure de test **on contrôle le risque** α mais pas le risque β. En d'autres termes, dans un test, on souhaite avant tout limiter à un faible niveau le risque de rejeter à tort la spécification H_0, se souciant moins d'accepter à tort, quand H_1 est vraie, cette même spécification. On peut encore dire que le rejet d'une hypothèse nulle est une véritable décision alors que son acceptation est plutôt un défaut de rejet. Face, par exemple, à une spécification sur une caractéristique d'un produit, le fait de rejeter cette spécification est une preuve quasi irréfutable qu'elle n'est pas correcte, alors que le fait de l'accepter ne signifie pas qu'elle soit correcte mais simplement que, sur la base des observations effectuées, rien ne permet de conclure qu'elle soit fausse. Notons que H_0 et H_1 ne sont pas interchangeables car la construction du test, via le choix de α, repose sur H_0 et non pas sur H_1. En particulier il n'est pas nécessaire de connaître la loi de T sous H_1, ce qui est d'ailleurs le cas pour la plupart des tests, y compris parmi les plus usuels.

Nous en venons maintenant à préciser cette idée de «pertinence» de la statistique, tant il est vrai qu'il ne suffit évidemment pas de choisir n'importe quelle statistique de loi connue sous H_0. Il est naturel de poser comme exigence que la statistique ait une plus forte propension à tomber dans la région de rejet quand H_1 est la bonne hypothèse, ce que nous transcrivons mathématiquement par la condition que la probabilité de rejeter H_0 soit plus élevée sous H_1 que sous H_0, et si possible nettement plus élevée. Toute la recherche, intuitive ou non, d'une bonne statistique de test repose sur ce principe que nous allons maintenant formaliser avec la notion de puissance.

Définition 9.3 *On appelle* ***puissance d'un test*** *la probabilité de rejeter H_0 alors qu'elle est effectivement fausse soit, dans les notations précédentes :*

$$P(T \in \overline{A} \,|\, H_1)\,.$$

La puissance, qui est la capacité à détecter qu'une hypothèse nulle est fausse, n'est rien d'autre que $1-\beta$ puisque β, le risque de première espèce, est la probabilité de l'événement complémentaire également sous H_1. Nous pouvons maintenant clairement exprimer notre exigence.

Définition 9.4 *On dit qu'un test est* ***sans biais*** *si sa puissance est supérieure ou égale à son risque α, soit :*

$$P(T \in \overline{A} \,|\, H_1) \geq P(T \in \overline{A} \,|\, H_0).$$

En conclusion une condition naturelle pour qu'une statistique soit éligible pour tester une hypothèse est qu'elle induise un test sans biais. Incidemment ce terme de «sans biais» n'a pas de rapport direct avec la notion de biais d'un estimateur.

On entrevoit dès lors que le choix entre plusieurs tests potentiels, pour une hypothèse nulle donnée, se jouera sur la puissance. Avant de préciser cela notons qu'un test, tel que nous avons présenté les choses, est parfaitement défini par le couple : statistique de test et région d'acceptation (T, A), puisque α, β et la puissance $1-\beta$ en découlent (même si conceptuellement le choix de α précède celui de A, mais pour α fixé il y a différentes façons de choisir une région - généralement un intervalle - de probabilité α sur la loi de T sous H_0).

En vérité il n'est pas nécessaire de se référer à une statistique de test. En effet mettons en évidence la fonction de l'échantillon définissant la statistique : $T = h(X_1, X_2, \cdots, X_n)$ et soit $\mathbb{A}$ l'ensemble des points de $\mathbb{R}^n$, réalisations de $(X_1, X_2, \cdots, X_n)$, défini par :

$$\mathbb{A} = \{(x_1, x_2, \cdots, x_n) \,|\, h(x_1, x_2, \cdots, x_n) \in A\}\,.$$

L'événement $(T \in A)$, que ce soit sous H_0 ou sous H_1, est identique à l'événement $((X_1, X_2, \cdots, X_n) \in \mathbb{A})$. Le test est donc parfaitement défini par la région d'acceptation $\mathbb{A}$ dans $\mathbb{R}^n$. D'une façon générale **un test s'identifie à une région d'acceptation dans l'espace des réalisations**. Cette vision plus fondamentale sera parfois utile dans les développements à venir, bien que ce ne soit qu'une vue de l'esprit dans la mesure où une règle de décision fondée sur une région dans un espace à n dimensions n'est pas praticable et que tout test, ou presque, passe par une statistique à valeurs dans $\mathbb{R}$ avec une région d'acceptation sous forme d'un intervalle. Ayant dégagé la définition d'un test nous pouvons aborder la comparaison de divers tests.

Définition 9.5 *On dit que le test τ_1 est **plus puissant** que le test τ_2 au niveau α s'il est de niveau α, si τ_2 est de niveau égal (ou inférieur) à α et si la puissance de τ_1 est supérieure à celle de τ_2.*

Il est évident que toute comparaison de puissance doit s'opérer à un même niveau. En effet pour tout test il y a un lien entre risque α et puissance : en prenant un risque α plus élevé on agrandit la région de rejet $\overline{A}$ et, par voie de conséquence, on augmente également la puissance. Notons aussi que le fait de comparer τ_1 à τ_2 qui serait à un risque α plus faible est pénalisant pour ce dernier, mais cette éventualité aura sa raison d'être, notamment dans le cas discret.

L'objectif sera finalement de rechercher **le test le plus puissant** parmi tous. Dans le cas où H_0 et H_1 sont des hypothèses simples il existe un tel test, mais cela n'est pas nécessairement vrai dans le cas où l'hypothèse alternative

est multiple. Par ailleurs, en général, quand une statistique de test donne le test le plus puissant à un niveau donné elle reste optimale à tout autre niveau.

Remarques

1. Dans le cas d'une loi discrète la statistique sera elle-même discrète et le niveau α choisi ne pourra être exactement atteint. Comme pour les intervalles de confiance, si l'on souhaite un risque de première espèce de 0,05, par exemple, on recherchera une région $\overline{A}$ de probabilité, sous H_0, la plus proche possible mais inférieure à 0,05. On dira alors que l'on a un *test conservateur*. Ceci justifie, au demeurant, la comparaison de τ_2 à τ_1 selon la définition 9.5 à un niveau de τ_2 éventuellement inférieur à celui de τ_1.
2. Les définitions ci-dessus s'appliquent à des situations d'hypothèses multiples (et même non paramétriques) moyennant quelques précisions que nous donnerons en temps utile.
3. Nous développons ici la théorie des tests telle qu'elle a été formalisée par J. Neyman et E.S. Pearson autour de 1930. La pratique s'est aujourd'hui éloignée de la théorie. En particulier, le choix a priori d'un niveau α ne correspond pas à l'usage, sauf dans des protocoles de tests définis par exemple par une réglementation (notamment les tests pharmaceutiques). Il n'empêche que le cadre théorique classique reste indispensable pour élaborer les bonnes méthodes.
4. Il est une autre exigence, outre celle de «sans biais», que l'on doit avoir pour une bonne procédure de test, à savoir que lorsque la taille de l'échantillon tend vers l'infini, la suite de tests correspondante $\{\tau_n\}$ soit telle que la puissance β_n s'accroisse et tende vers 1. En d'autres termes on doit avoir la garantie que l'on gagne à observer de très grands échantillons, étant pratiquement sûr, à la limite, de détecter une hypothèse nulle qui serait fausse. On dit alors que la procédure de test est *convergente.*

Donnons deux exemples, certes quelque peu artificiels, mais illustrant les notions introduites, l'un dans le cas continu, l'autre dans le cas discret.

Exemple 9.1 Supposons que deux machines A et B produisent le même type de produit, mais la machine A fournit un produit plus cher de qualité supérieure. La qualité d'un produit se mesure à une entité aléatoire qui est de loi $\mathcal{N}(5\,;1)$ pour la machine A et $\mathcal{N}(4\,;1)$ pour la machine B, et ne diffère donc que par la moyenne. Un client achète le produit le plus cher par lots de 10 et désire développer un test pour contrôler qu'un lot donné provient bien de la machine A. Comme accuser le producteur à tort peut avoir de graves conséquences, il doit limiter le risque correspondant et tester $H_0 : \mu = 5$ vs. $H_1 : \mu = 4$, à un niveau 0,05 par exemple. Il semble naturel d'utiliser comme statistique de test la moyenne $\overline{X}$ du lot. Sous H_0 sa loi est $\mathcal{N}(5\,;1/10)$ et l'on a alors l'intervalle

de probabilité 0,95 : $[5-1{,}96/\sqrt{10}\,;\ 5+1{,}96/\sqrt{10}]$, soit [4,38 ; 5,62]. D'où une règle de décision simple :

- accepter H_0 si la réalisation $\overline{x}$ (moyenne du lot considéré) de $\overline{X}$ est dans [4,38 ; 5,62],
- rejeter sinon.

Il est possible de calculer la puissance de ce test puisque la loi de $\overline{X}$ est connue sous H_1 : c'est la loi $\mathcal{N}(4\,;1/10)$. Le risque de deuxième espèce vaut :

$$\begin{aligned}\beta &= P(4{,}38 < \overline{X} < 5{,}62 \,|\, H_1)\\ &= P(\frac{4{,}38-4}{1/\sqrt{10}} < Z < \frac{5{,}62-4}{1/\sqrt{10}}) \quad \text{avec } Z \rightsquigarrow \mathcal{N}(0\,;1)\\ &= P(1{,}20 < Z < 5{,}12) \simeq 0{,}115.\end{aligned}$$

D'où une puissance d'environ 0,885. Notons que l'on peut obtenir un test plus puissant en prenant comme région d'acceptation l'intervalle de probabilité 0,95 : $[5-1{,}645/\sqrt{10}\,;\ +\infty[$ où -1,645 est le quantile d'ordre 0,05 de la loi $\mathcal{N}(0\,;1)$, soit $[4{,}48\,;\ +\infty[$. En effet :

$$\beta = P(4{,}48 < \overline{X} \mid H_1) = P(\frac{4{,}48-4}{1/\sqrt{10}} < Z) = P(1{,}52 < Z) \simeq 0{,}064\,,$$

ce qui donne une puissance de 0,936. Intuitivement on sent bien que, dans le premier test, il est peu pertinent de borner la zone d'acceptation vers le haut car cela conduit à rejeter l'hypothèse nulle pour de très grandes valeurs de $\overline{x}$, au-delà de 5,62. ■

Exemple 9.2 On sait que le nombre de particules émises par une source radioactive par unité de temps suit une loi de Poisson. Observant l'émission d'un corps durant 20 unités de temps on doit décider s'il s'agit d'une source de type A versus une source de type B. La source A émet en moyenne 0,6 particules par unité de temps et la source B en émet 0,8. On teste donc $H_0 : \lambda = 0{,}6$ vs. $H_1 : \lambda = 0{,}8$. On peut construire un test sur la statistique $\sum_{i=1}^{20} X_i$, le nombre total de particules émises au cours des 20 unités de temps, qui suit une loi $\mathcal{P}(12)$ sous H_0. Intuitivement on choisit une région de rejet de la forme $\sum_{i=1}^{20} x_i \geq k$, puisqu'un nombre plutôt élevé de comptages va à l'encontre de l'hypothèse nulle. Choisissant a priori $\alpha = 0{,}05$, on lit dans une table (ou dans un logiciel) que pour une v.a. $T \rightsquigarrow \mathcal{P}(12)$ on a $P(T \geq 18) = 0{,}0630$ et $P(T \geq 19) = 0{,}0374$. On optera pour un test conservateur en rejetant H_0 si $\sum_{i=1}^{20} x_i \geq 19$.

La puissance du test est égale à $P(\sum_{i=1}^{20} X_i \geq 19 \mid H_1)$ soit, dans une table, $P(S \geq 19)$ où $S \rightsquigarrow \mathcal{P}(16)$. On trouve 0,258 qui montre que le test est sans biais. ■

Nous mettons maintenant en évidence le test le plus puissant.

9.3 Test du rapport de vraisemblance simple

9.3.1 Propriété d'optimalité

Reprenons la fonction de vraisemblance du paramètre inconnu θ (voir définition 6.11) pour une réalisation de l'échantillon $(x_1, x_2, \cdots, x_n)$:

$$L(\theta; x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Nous restons dans le cadre d'une hypothèse nulle et d'une hypothèse alternative simples, soit $\Theta = \{\theta_0, \theta_1\}$. On supposera dans cette section que le support de la densité $f(x; \theta)$ ne dépend pas de θ.

Définition 9.6 *On appelle* ***test du rapport de vraisemblance*** *(RV) de l'hypothèse $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ au niveau α, le test défini par la région de rejet de la forme :*

$$\frac{L(\theta_0; x_1, x_2, ..., x_n)}{L(\theta_1; x_1, x_2, ..., x_n)} < k_\alpha$$

où k_α est une valeur (positive) déterminée en fonction du risque de première espèce α.

Ce test a une certaine logique intuitive puisqu'il conduit à rejeter la valeur spécifiée θ_0 lorsqu'elle est moins vraisemblable que la valeur alternative θ_1, car k_α (dont on admettra l'existence) se trouvera, en fait, être plus petit que 1 pour garantir un risque α faible (voir exemple 9.3). Notons que le rapport des deux vraisemblances (que l'on nomme rapport **de** vraisemblance) est bien une statistique puisque θ_0 et θ_1 sont donnés.

Théorème 9.1 *(historiquement :* ***lemme de Neyman-Pearson****)*
Le test du RV est le plus puissant quel que soit le choix de $\alpha \in]0, 1[$.

Démonstration : soit $\mathbb{A}^* \subset \mathbb{R}^n$ la région d'acceptation associée au test du RV de niveau α, i.e. :

$$\mathbb{A}^* = \left\{ (x_1, x_2, ..., x_n) \mid \frac{L(\theta_0; x_1, x_2, ..., x_n)}{L(\theta_1; x_1, x_2, ..., x_n)} \geq k_\alpha \right\}$$

et $\mathbb{A}$ celle d'un quelconque autre test. Les risques de première espèce du test du RV et de l'autre test s'écrivent[2] donc, respectivement, $P(\overline{\mathbb{A}}^* \mid H_0)$ et $P(\overline{\mathbb{A}} \mid H_0)$

[2]Il n'est pas inutile de rappeler ici la convention initiale de la section 1.1, à savoir que pour une v.a. X quelconque $P(X \in A)$ est la probabilité $P(A)$ associée à la partie A de $\mathbb{R}$. Ici la probabilité $P(\overline{\mathbb{A}}^*)$, par exemple, aurait pu être notée $P((X_1, X_2, \cdots, X_n) \in \overline{\mathbb{A}}^*)$ pour se référer aux réalisations de l'échantillon dans $\mathbb{R}^n$. Même si cette notation est plus explicite nous y renonçons pour simplifier les écritures.

et, pour effectuer la comparaison, on doit avoir, en accord avec la définition 9.5, $P(\overline{\mathbb{A}}^* \,|\, H_0) - P(\overline{\mathbb{A}} \,|\, H_0) \geq 0$.

Partitionnons $\mathbb{A}^*$ selon $\mathbb{A}^* \cap \mathbb{A}$ et $\mathbb{A}^* \cap \overline{\mathbb{A}}$, et de même $\mathbb{A}$ selon $\mathbb{A} \cap \mathbb{A}^*$ et $\mathbb{A} \cap \overline{\mathbb{A}}^*$. La différence entre les risques de deuxième espèce des deux tests est :

$$P(\mathbb{A} \,|\, H_1) - P(\mathbb{A}^* \,|\, H_1) = P(\mathbb{A} \cap \overline{\mathbb{A}}^* \,|\, H_1) - P(\mathbb{A}^* \cap \overline{\mathbb{A}} \,|\, H_1)$$

puisque les probabilités sur la partie commune $\mathbb{A}^* \cap \mathbb{A}$ s'éliminent. Or en tout point de $\mathbb{A}^*$, et donc de $\mathbb{A}^* \cap \overline{\mathbb{A}}$, on a :

$$L(\theta_1; x_1, x_2, ..., x_n) \leq \frac{1}{k_\alpha} L(\theta_0; x_1, x_2, ..., x_n),$$
$$\text{i.e. } f(x_1, x_2, ..., x_n; \theta_1) \leq \frac{1}{k_\alpha} f(x_1, x_2, ..., x_n; \theta_0)$$

et, par intégration (ou sommation dans le cas discret) sur le domaine $\mathbb{A}^* \cap \overline{\mathbb{A}}$, on a donc :

$$P(\mathbb{A}^* \cap \overline{\mathbb{A}} \,|\, H_1) \leq \frac{1}{k_\alpha} P(\mathbb{A}^* \cap \overline{\mathbb{A}} \,|\, H_0).$$

Pour tout point de $\overline{\mathbb{A}}^*$, et donc de $\mathbb{A} \cap \overline{\mathbb{A}}^*$, l'inégalité s'inverse et on obtient :

$$P(\mathbb{A} \cap \overline{\mathbb{A}}^* \,|\, H_1) > \frac{1}{k_\alpha} P(\mathbb{A} \cap \overline{\mathbb{A}}^* \,|\, H_0).$$

En revenant à l'équation de départ, il s'ensuit que :

$$P(\mathbb{A} \,|\, H_1) - P(\mathbb{A}^* \,|\, H_1) > \frac{1}{k_\alpha} \left[P(\mathbb{A} \cap \overline{\mathbb{A}}^* \,|\, H_0) - P(\mathbb{A}^* \cap \overline{\mathbb{A}} \,|\, H_0) \right].$$

Considérant les partitions indiquées plus haut, le terme de droite est aussi égal à $\frac{1}{k_\alpha} [P(\mathbb{A} \,|\, H_0) - P(\mathbb{A}^* \,|\, H_0)]$, lequel est lui-même égal à :

$$\frac{1}{k_\alpha} \left[1 - P(\overline{\mathbb{A}} \,|\, H_0) - 1 + P(\overline{\mathbb{A}}^* \,|\, H_0) \right] = \frac{1}{k_\alpha} \left[P(\overline{\mathbb{A}}^* \,|\, H_0) - P(\overline{\mathbb{A}} \,|\, H_0) \right]$$

qui, par hypothèse, est positif ou nul. Ainsi le risque β d'un test quelconque est strictement supérieur à celui du test du RV et, de façon équivalente, il est donc moins puissant. □

Note 9.1 Dans le cas discret la démonstration du théorème peut poser problème. En effet on n'a pas nécessairement $P(\overline{\mathbb{A}}^* \,|\, H_0) - P(\overline{\mathbb{A}} \,|\, H_0) \geq 0$ dans la mesure où le risque de première espèce réel du test du RV peut, par conservatisme, être inférieur au niveau nominal α, alors que celui de l'autre test peut être plus proche de α. En fait le même résultat d'optimalité peut être démontré à condition de «randomiser» la règle de décision (test randomisé ou test mixte).

Soit k_a la valeur qui donne une probabilité de rejet sous H_0 aussi proche que possible du niveau nominal α mais inférieure (exceptionnellement égale) à α, $k_a + 1$

donnant alors une probabilité supérieure à α. La randomisation consiste à choisir la limite k_α avec une probabilité p et $k_a + 1$ avec probabilité $1 - p$. On doit déterminer p pour que finalement le risque résultant soit exactement α. Appliqué à l'exemple 9.2 ce procédé conduirait à rejeter selon la règle $\sum_{i=1}^{20} x_i \geq 19$ avec probabilité p et selon $\sum_{i=1}^{20} x_i \geq 18$ avec probabilité $1 - p$, la valeur de p étant telle que $p \times 0{,}0374 + (1 - p) \times 0{,}0630 = 0{,}05$ soit $p = 0{,}51$. A peu de choses près on doit jouer à pile ou face le choix de la règle avec 19 ou celui de la règle avec 18.

On peut aussi ajuster son choix de α en prenant une probabilité exactement atteinte (dans l'exemple ci-dessus on pourra prendre la règle $\sum_{i=1}^{20} x_i \geq 19$ en fixant $\alpha = 0,0374$). Alors la démonstration ci-dessus est valide.

Note 9.2 On a dû supposer que la densité ait un support indépendant de θ pour éviter que $L(\theta_1; x_1, x_2, ..., x_n)$ s'annule alors que $L(\theta_0; x_1, x_2, ..., x_n)$ ne s'annule pas. On peut toutefois contourner ce problème en définissant la région de rejet par $L(\theta_0; x_1, x_2, ..., x_n) < k\, L(\theta_1; x_1, x_2, ..., x_n)$. Si $L(\theta_1; x_1, x_2, ..., x_n)$ s'annule on a une réalisation impossible sous H_1 et l'on peut choisir θ_0 sans aucun risque d'erreur.

Proposition 9.1 *Le test du RV est sans biais.*

Démonstration : pour rester très général ne supposons pas que k_α soit inférieur à 1. Si $k_\alpha \geq 1$ on a, pour tout point de la région d'acceptation $\mathbb{A}^*$:

$$f(x_1, x_2, ..., x_n; \theta_0) \geq f(x_1, x_2, ..., x_n; \theta_1)$$

et, par conséquent, $P(\mathbb{A}^* \mid H_0) \geq P(\mathbb{A}^* \mid H_1)$ d'où $P(\overline{\mathbb{A}}^* \mid H_0) \leq P(\overline{\mathbb{A}}^* \mid H_1)$ qui est la condition requise. Inversement, si $k_\alpha < 1$, on a, pour tout point de la région de rejet $\overline{\mathbb{A}}^*$:

$$f(x_1, x_2, ..., x_n; \theta_0) < f(x_1, x_2, ..., x_n; \theta_1)$$

et, par conséquent, il est vrai aussi que $P(\overline{\mathbb{A}}^* \mid H_0) < P(\overline{\mathbb{A}}^* \mid H_1)$. □

Moyennant des conditions mineures on peut également démontrer que la procédure du RV est convergente.

Ayant mis en évidence le test le plus puissant se pose la question de la faisabilité de ce test. En effet pour qu'il puisse être mis en oeuvre il faut que la statistique du rapport de vraisemblance prenne une forme telle que sa loi soit connue sous H_0. Montrons sur l'exemple 9.1 que le test du RV peut se ramener à une forme simple ce que nous démontrerons ensuite pour les lois de la classe exponentielle.

Exemple 9.3 Dans le contexte de l'exemple 9.1 la fonction de vraisemblance pour μ est :

$$\begin{aligned} L(\mu; x_1, x_2, \cdots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x_i - \mu)^2\} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\}. \end{aligned}$$

Pour être plus général considérons $H_0 : \mu = \mu_0$ vs. $H_1 : \mu = \mu_1$ d'où le rapport de vraisemblance :

$$\begin{aligned} \frac{L(\mu_0; x_1, x_2, \cdots, x_n)}{L(\mu_1; x_1, x_2, \cdots, x_n)} &= \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \mu_1)^2\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[-2(\mu_0 - \mu_1)\sum_{i=1}^{n} x_i + n(\mu_0^2 - \mu_1^2)\right]\right\}. \end{aligned}$$

Le rapport de vraisemblance ne dépend donc des observations qu'à travers $\sum_{i=1}^{n} x_i$. De plus comme c'est une fonction croissante de $(\mu_0 - \mu_1)\sum_{i=1}^{n} x_i$ il est inférieur à k_α si et seulement si $(\mu_0 - \mu_1)\sum_{i=1}^{n} x_i < k'_\alpha$ où k'_α est une autre constante que l'on déduit aisément de k_α. Si $\mu_0 > \mu_1$ alors la région de rejet est de la forme $\sum_{i=1}^{n} x_i < k''_\alpha$ ou, de façon équivalente, $\overline{x} < k'''_\alpha$. Si $\mu_0 < \mu_1$ les inégalités doivent être inversées.

Dans l'exemple 9.1 on a vu que, pour $\alpha = 0{,}05$, $\mu_0 = 5$ et $\mu_1 = 4$, la région de rejet était définie par $\overline{x}$ <4,48 pour le deuxième test envisagé, lequel s'avère être le test le plus puissant.

Ainsi, bien que la statistique du rapport de vraisemblance elle-même ne soit pas simple, le fait qu'elle soit **fonction monotone** de $\sum_{i=1}^{n} X_i$ dont la loi est connue suffit pour mettre au point le test. Par curiosité calculons la constante k_α propre au rapport de vraisemblance. Comme $n = 10$ on a $k''_\alpha = 44{,}8 = k'_\alpha$ puisque $\mu_0 - \mu_1 = 1$. Alors :

$$k_\alpha = \exp\left\{-\frac{1}{2}\left[-2 \times 44{,}8 + 10(25 - 16)\right]\right\} = 0{,}82.$$

Ceci signifie que le test du RV consiste à rejeter $H_0 : \mu = 5$ vs. $H_1 : \mu = 4$ lorsque la vraisemblance de la valeur 5 du paramètre inconnu μ est 0,82 fois celle de la valeur 4.

Notons encore que si l'on avait eu $\mu_0 < \mu_1$ la région de rejet aurait été de la forme $\overline{x} > c$, ce qui correspond à l'intuition. ■

9.3.2 Cas d'un paramètre de dimension 1

Nous montrons ici que le test du RV se ramène à un test simple, comme dans l'exemple ci-dessus, dans une grande variété de situations.

Proposition 9.2 *S'il existe une statistique $T = t(X_1, X_2, \cdots, X_n)$ exhaustive minimale à valeurs dans $\mathbb{R}$ alors le rapport de vraisemblance ne dépend de la réalisation $(x_1, x_2, ..., x_n)$ qu'à travers la valeur $t(x_1, x_2, ..., x_n)$. De plus si ce rapport de vraisemblance est une fonction monotone de $t(x_1, x_2, ..., x_n)$ alors le test du RV se ramène à un test dont la région de rejet est de la forme $t(x_1, x_2, ..., x_n) < c$ si c'est une fonction croissante ou $t(x_1, x_2, ..., x_n) > c$ si la fonction est décroissante.*

En effet, par le théorème de factorisation 6.1, si $T = t(X_1, X_2, \cdots, X_n)$ est une statistique exhaustive minimale alors l'expression de la vraisemblance $L(\theta; x_1, x_2, \cdots, x_n)$ qui est identique à l'expression de la densité conjointe est de la forme $g(t(x_1, x_2, ..., x_n); \theta)\, h(x_1, x_2, ..., x_n)$ et le RV ne dépend plus que du rapport $g(t(x_1, x_2, ..., x_n); \theta_0) \,/\, g(t(x_1, x_2, ..., x_n); \theta_1)$. Si $g(u; \theta_0) \,/\, g(u; \theta_1)$ est une fonction monotone de u alors l'inégalité $g(u; \theta_0) \,/\, g(u; \theta_1) < k$ est équivalente à une inégalité sur u.

Proposition 9.3 *Si la loi mère est dans une famille appartenant à la classe exponentielle, i.e. $f(x; \theta) = a(\theta)\, b(x)\, \exp\{c(\theta)\, d(x)\}$ (voir section 6.3), alors le test du RV a une région de rejet de la forme :*

$$\sum_{i=1}^{n} d(x_i) < k \quad \textit{si } c(\theta_0) - c(\theta_1) > 0$$

$$\textit{ou} \quad \sum_{i=1}^{n} d(x_i) > k \quad \textit{si } c(\theta_0) - c(\theta_1) < 0.$$

Cette proposition est un corollaire de la précédente puisque l'on a vu que $\sum_{i=1}^{n} d(X_i)$ est exhaustive minimale (voir proposition 6.5). Le RV étant égal à $[a(\theta_0)/a(\theta_1)]^n \exp\{[c(\theta_0) - c(\theta_1)] \sum_{i=1}^{n} d(x_i)\}$ on voit que le sens de l'inégalité dépendra du signe de $c(\theta_0) - c(\theta_1)$. Notons que très souvent la loi de $\sum_{i=1}^{n} d(X_i)$ est de type connu et k sera donc le quantile d'ordre α de cette loi sous H_0 ou d'ordre $1-\alpha$, selon que le signe est positif ou négatif. On pourra également calculer la puissance de ce test optimal. Les exemples 9.1 et 9.2 (loi de Gauss de variance connue et loi de Poisson) correspondaient à cette situation. Considérons encore le cas d'une loi de Bernoulli.

Exemple 9.4 Soit le test $H_0 : p = p_0$ vs. $H_1 : p = p_1$ pour le paramètre p d'une loi de Bernoulli. Cette famille de lois appartient à la classe exponentielle et l'on a : $f(x; p) = p^x (1-p)^{1-x} = \exp\{\ln(\frac{p}{1-p})\, x\}$ pour $x \in \{0, 1\}$. Donc $d(x) = x$ et, supposant par exemple que $p_0 > p_1$, on a $\ln \frac{p_0}{1-p_0} > \ln \frac{p_1}{1-p_1}$ et la région de rejet est de la forme $\sum_{i=1}^{n} x_i < k$ ou préférablement, comme nous sommes

dans le cas discret, $\sum_{i=1}^n x_i \leq k'$. La statistique de test $\sum_{i=1}^n X_i$ suit une loi $\mathcal{B}(n,p)$ et pour $\alpha = 0{,}05$, k' est la valeur égale ou immédiatement inférieure au quantile d'ordre 0,05 sur la loi $\mathcal{B}(n,p_0)$. La puissance est la probabilité d'être inférieur ou égal à k' pour la loi $\mathcal{B}(n,p_1)$.

Soit par exemple $H_0 : p = 0{,}5$ vs. $H_1 : p = 0{,}3$ à tester au niveau 0,05 avec un échantillon de taille 30. On lit dans une table binomiale pour la loi $\mathcal{B}(30\,;0{,}5)$: $P(\sum_{i=1}^n x_i \leq 10) = 0{,}0494$ et $P(\sum_{i=1}^n x_i \leq 11) = 0{,}1003$. On choisit donc la règle de rejet $\sum_{i=1}^n x_i \leq 10$. Pour la puissance on lit sur la loi $\mathcal{B}(30\,;0{,}3)$: $P(\sum_{i=1}^n x_i \leq 10) = 0{,}730$. ■

Le cas de deux hypothèses simples, nous l'avons dit, est peu réaliste et il nous faut maintenant envisager des situations plus générales.

9.4 Tests d'hypothèses multiples

9.4.1 Risques, puissance et optimalité

Lorsque l'une des hypothèses H_0 ou H_1 est multiple les définitions de la section 9.2 doivent être revues. En effet si dans une hypothèse plusieurs valeurs du paramètre sont possibles il n'y a plus de risque unique. Ainsi une expression telle que $P(T \in \overline{A}\,|\,H_0)$ n'a pas de sens si H_0 est multiple.

Plaçons-nous dans le cas le plus général où l'on souhaite tester :

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta_1,$$

où $\Theta_1 = \Theta - \Theta_0$ est le complémentaire de Θ_0 par rapport à Θ. Comme précédemment, de la façon la plus générale, un test est défini par une région $\mathbb{A} \subset \mathbb{R}^n$ d'acceptation de l'hypothèse nulle H_0. Nous supposerons ici, comme cela se trouve en pratique, que cette région se réduit à un intervalle A de $\mathbb{R}$ pour une statistique de test T. Alors la règle de décision consiste à accepter H_0 si la valeur réalisée t de T appartient à A et à rejeter H_0 sinon. Si H_0 est multiple le risque de première espèce $P_\theta(T \in \overline{A})$ dépend de θ appartenant à Θ_0. Le niveau du test est alors défini comme le risque maximal que l'on encourt à rejeter H_0 alors qu'elle serait fausse.

Définition 9.7 *Soit $H_0 : \theta \in \Theta_0$ une hypothèse nulle multiple et $\alpha(\theta)$ le risque de première espèce pour la valeur $\theta \in \Theta_0$. On appelle* ***niveau du test*** *(ou seuil du test) la valeur α telle que :*

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)\,.$$

De même si l'hypothèse alternative $H_1 : \theta \in \Theta_1$ est multiple le risque de deuxième espèce est une fonction $\beta(\theta)$ ainsi que la puissance. On définit alors la **fonction puissance** du test :

$$h(\theta) = 1 - \beta(\theta) = P_\theta(T \in \overline{A}) \text{ définie pour tout } \theta \in \Theta_1.$$

Définition 9.8 *On dit qu'un test est* ***sans biais*** *si sa fonction puissance reste supérieure ou égale à son niveau α, soit :*

$$P_\theta(T \in \overline{A}) \geq \alpha \quad \textit{pour tout } \theta \in \Theta_1.$$

En d'autres termes, la probabilité de rejeter H_0 si elle est fausse, quelle que soit la valeur de θ dans Θ_1, est toujours plus élevée que la probabilité de la rejeter si elle est vraie, quelle que soit alors la valeur de θ dans Θ_0.

Définition 9.9 *On dit que le test τ_1 de niveau α est* ***uniformément plus puissant*** *que le test τ_2 au niveau α s'il est de niveau α, si τ_2 est de niveau égal (ou inférieur) à α et si la fonction puissance de τ_1 reste toujours supérieure ou égale à celle de τ_2, mais strictement supérieure pour au moins une valeur de $\theta \in \Theta_1$, i.e. pour tout $\theta \in \Theta_1$, $h_1(\theta) \geq h_2(\theta)$ et il existe $\theta^* \in \Theta_1$ tel que $h_1(\theta^*) > h_2(\theta^*)$, où $h_1(\theta)$ et $h_2(\theta)$ sont les fonctions puissance respectives des tests τ_1 et τ_2.*

Le terme «uniformément» se rapporte au fait que la puissance de τ_1 est supérieure quelle que soit $\theta \in \Theta_1$.

Définition 9.10 *On dit que le test τ^* est* ***uniformément le plus puissant*** *(UPP) au niveau α s'il est uniformément plus puissant que tout autre test au niveau α.*

Rien ne dit qu'un tel test existe. En effet il se peut, par exemple, qu'un premier test domine tous les autres pour certaines valeurs de θ dans Θ_1, qu'un deuxième soit le meilleur pour d'autres valeurs, etc. Signalons que certains ouvrages en français parlent de test UMP (de l'anglais *uniformly most powerful*).

Dans la situation la plus générale il n'existera généralement pas de test UPP. Néanmoins le résultat de Neyman-Pearson obtenu dans la situation simple de la section 9.3 s'étend assez naturellement à des situations d'hypothèses multiples dites unilatérales, très fréquentes en pratique.

9.4.2 Tests d'hypothèses multiples unilatérales

Nous considérons dans cette section des situations de test du type :

$$\begin{aligned} & H_0 : \theta \leq \theta_0 \qquad \text{vs.} \qquad H_1 : \theta > \theta_0 \\ \text{ou} \quad & H_0 : \theta \geq \theta_0 \qquad \text{vs.} \qquad H_1 : \theta < \theta_0 \end{aligned}$$

où θ est donc un paramètre de dimension 1 ($\Theta \subseteq \mathbb{R}$). Hypothèse nulle et hypothèse alternative sont multiples. De telles situations se rencontrent lorsque l'on s'intéresse uniquement à juger si le paramètre θ dépasse un certain seuil (par exemple une norme de qualité, un seuil de pollution, un niveau antérieur, etc.). L'hypothèse nulle est dite *unilatérale* et par extension on parle aussi de *test unilatéral* du fait que la région de rejet est usuellement de la forme $T > c$ ou $T < c$, T étant la statistique de test.

Proposition 9.4 *S'il existe une statistique $T = t(X_1, X_2, \cdots, X_n)$ exhaustive minimale à valeurs dans $\mathbb{R}$ et si, pour tout couple (θ, θ') tel que $\theta < \theta'$, le rapport de vraisemblance $L(\theta; x_1, x_2, ..., x_n)/L(\theta'; x_1, x_2, ..., x_n)$ est une fonction monotone de $t(x_1, x_2, \cdots, x_n)$, alors il existe un test uniformément le plus puissant pour les situations d'hypothèses unilatérales et la région de rejet est soit de la forme $t(x_1, x_2, \cdots, x_n) < k$, soit de la forme $t(x_1, x_2, \cdots, x_n) > k$.*

En proposition 9.2 on a vu que le RV ne pouvait dépendre que de $t(x_1, x_2, \cdots, x_n)$ et que, dans le cas où l'hypothèse nulle et l'hypothèse alternative sont simples, il suffisait, pour se ramener à une région de rejet de la forme $t(x_1, x_2, \cdots, x_n) < k$ ou $t(x_1, x_2, \cdots, x_n) > k$, que le RV soit monotone pour les deux valeurs de θ concernées, l'existence d'un test le plus puissant étant quoi qu'il en soit assurée. Ici la monotonicité du RV pour tout couple (θ, θ') est requise afin, d'une part, de garantir l'existence d'un test UPP et, d'autre part, de ramener ce test à une simple inégalité sur $t(x_1, x_2, \cdots, x_n)$.

Démonstration de la proposition : prenons le cas où le RV est une fonction croissante de $t(x_1, x_2, \cdots, x_n)$ et pour $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$.

Montrons tout d'abord que $P_\theta(T < k)$ croît avec θ. Considérons le test simple fictif : $H_0' : \theta = \theta'$ vs. $H_1' : \theta = \theta''$ avec $\theta' < \theta''$. Le test à région de rejet $t(x_1, x_2, \cdots, x_n) < k$ est équivalent au test du RV avec

$$L(\theta'; x_1, x_2, ..., x_n)/L(\theta''; x_1, x_2, ..., x_n) < k'$$

et est donc sans biais (voir proposition 9.1), d'où $P_{\theta'}(T < k) \leq P_{\theta''}(T < k)$ quels que soient θ' et θ'' tels que $\theta' < \theta''$.

Supposons maintenant que nous choisissions $t(x_1, x_2, \cdots, x_n) < k$ comme région de rejet pour tester $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Alors $P_\theta(T < k)$ correspond à la probabilité de rejet pour une valeur quelconque θ. En particulier le risque de première espèce croît pour $\theta \in]-\infty\ ,\ \theta_0]$ et le risque maximal est donc atteint en θ_0. Par conséquent, pour obtenir un niveau α il suffit de choisir k tel que $P_{\theta_0}(T < k) = \alpha$. On notera au passage que $P_\theta(T < k)$, pour $\theta \in]\theta_0\ ,+\infty[$, définit la fonction puissance laquelle est également croissante, et ceci au fur et à mesure que l'on s'éloigne de θ_0.

Du fait que le risque de première espèce est maximal en θ_0, on peut se contenter d'étudier la situation de test restreinte $\widetilde{H}_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$. Soit θ' une valeur dans H_1, alors $t(x_1, x_2, \cdots, x_n) < k$ équivaut à :

$$L(\theta_0; x_1, x_2, ..., x_n)/L(\theta'; x_1, x_2, ..., x_n) < k',$$

qui correspond au test du RV simple de puissance $h(\theta')$ supérieure à celle de tout autre test en vertu du théorème 9.1. Ceci restant vrai quel que soit θ' dans H_1, le test $t(x_1, x_2, \cdots, x_n) < k$ est bien uniformément le plus puissant. □

La démonstration a été faite dans un cas particulier, mais elle est analogue pour les trois autres cas possibles. Pour la situation $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$

l'inégalité sur $t(x_1, x_2, \cdots, x_n)$ doit être inversée par rapport à la situation précédente car la région de rejet $L(\theta_0; x_1, x_2, ..., x_n)/L(\theta'; x_1, x_2, ..., x_n) < k'$ est alors définie avec une valeur θ_0 supérieure à θ' et le RV change donc de sens de variation par rapport à $t(x_1, x_2, \cdots, x_n)$. En résumé on a les régions de rejet suivantes :

1. $H_0 : \theta \leq \theta_0$ et RV fonction croissante de $t(x_1, \cdots, x_n) : t(x_1, \cdots, x_n) < k$
2. $H_0 : \theta \leq \theta_0$ et RV fonction décroissante de $t(x_1, \cdots, x_n) : t(x_1, \cdots, x_n) > k$
3. $H_0 : \theta \geq \theta_0$ et RV fonction croissante de $t(x_1, \cdots, x_n) : t(x_1, \cdots, x_n) > k$
4. $H_0 : \theta \geq \theta_0$ et RV fonction décroissante de $t(x_1, \cdots, x_n) : t(x_1, \cdots, x_n) < k$.

Notons que la propriété établie en préambule de la démonstration s'étend aux trois autres cas : **la probabilité de rejet est une fonction monotone du paramètre θ, le risque de première espèce est maximal en θ_0, la fonction puissance croît quand on s'éloigne de θ_0** (voir illustration de l'exemple 9.5 et figure 9.1).

Certains auteurs considèrent la situation $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$. On a vu au cours de la démonstration que celle-ci est équivalente à la situation $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ pour le test UPP.

Une famille de densité qui vérifie les conditions de la proposition 9.4 est dite être à **rapport de vraisemblance monotone.** Si elle appartient à la classe exponentielle son RV est égal à $[a(\theta)/a(\theta')]^n \exp\{[c(\theta) - c(\theta')] \sum_{i=1}^n d(x_i)\}$ (voir démonstration de la proposition 9.3). Dans cette classe une condition nécessaire et suffisante pour remplir ces conditions est donc que $c(\theta) - c(\theta')$ garde le même signe quel que soit le couple (θ, θ') tel que $\theta < \theta'$, c'est-à-dire que la fonction $c(\theta)$ soit monotone. De plus la fonction des observations $t(x_1, x_2, \cdots, x_n)$ de la proposition 9.4 est $\sum_{i=1}^n d(x_i)$, d'où la proposition suivante.

Proposition 9.5 *Si la loi mère est dans une famille appartenant à la classe exponentielle, i.e.* $f(x; \theta) = a(\theta)b(x)exp\{c(\theta)d(x)\}$, *et si la fonction* $c(\theta)$ *est monotone, alors il existe un test uniformément le plus puissant pour les situations d'hypothèses unilatérales et la région de rejet est soit de la forme* $\sum_{i=1}^n d(x_i) < k$, *soit de la forme* $\sum_{i=1}^n d(x_i) > k$.

Exemple 9.5 Comme dans les exemples 9.1 et 9.3 considérons une loi mère $\mathcal{N}(\mu\ ,\ 1)$ où μ est inconnu. Testons $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$. La densité de la loi mère est :

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

Ici $c(\mu) = \mu$ et $d(x) = x$. Le test UPP a donc une région de rejet de la forme $\sum_{i=1}^n x_i > k$ ou $\sum_{i=1}^n x_i < k$. Pour trouver le sens correct de l'inégalité il faut repartir du rapport de vraisemblance $L(\mu; x_1, x_2, ..., x_n)/L(\mu'; x_1, x_2, ..., x_n)$ qui,

comme il a été établi dans l'exemple 9.3, varie comme $\exp\{(\mu-\mu')\sum_{i=1}^n x_i\}$ et, pour $\mu<\mu'$, est donc une fonction décroissante de $\sum_{i=1}^n x_i$. Ainsi un RV inférieur à k est équivalent à $\sum_{i=1}^n x_i > k'$ ou encore à $\overline{x} > k''$. Ceci définit la région de rejet puisque nous sommes dans le cas 2 ci-dessus. Notons qu'on pouvait trouver intuitivement le sens de l'inégalité du fait qu'une moyenne empirique élevée abonde dans le sens de H_1 à l'encontre de H_0.

Dans l'hypothèse nulle on sait que le risque de première espèce est le plus élevé en μ_0. Pour cette valeur du paramètre μ, $\overline{X}$ est de loi $\mathcal{N}(\mu_0, 1/\sqrt{n})$. Pour un niveau 0,05 la constante k'' est définie par $P_{\mu_0}(\overline{X} > k'') = 0{,}05$ et est donc égale au quantile 0,95 de la loi $\mathcal{N}(\mu_0, 1/\sqrt{n})$, soit μ_0+1,645$(1/\sqrt{n})$. La fonction puissance est définie par $h(\mu) = P_\mu(\overline{X} > \mu_0$+1,645$(1/\sqrt{n}))$ pour $\overline{X} \rightsquigarrow \mathcal{N}(\mu, 1/\sqrt{n})$ et $\mu \in]\mu_0, +\infty[$. En posant $Z = \sqrt{n}(\overline{X}-\mu)$ on obtient :

$$h(\mu) = P(Z > 1{,}645 - \sqrt{n}(\mu-\mu_0)\,)$$

où $Z \rightsquigarrow \mathcal{N}(0\,;1)$. Comme $\mu-\mu_0 > 0$ la valeur 1,645$-\sqrt{n}(\mu-\mu_0)$ tend vers $-\infty$ quand n tend vers l'infini, donc $h(\mu)$ tend vers 1, ce qui démontre la convergence du test.

Pour illustrer cela considérons un produit dont une certaine mesure de qualité est, selon le producteur, inférieure ou égale à 5 en moyenne (par exemple la teneur en lipides d'un aliment allégé) et soit un test s'appuyant sur échantillon de taille 10. On considère toujours que l'écart-type de la variable qualité est connu et égal à 1. Pour un niveau de test 0,05 on rejettera H_0 si la teneur moyenne observée sur les 10 produits (tirés au hasard) est supérieure ou égale à 5+1,645$(1/\sqrt{10})$ = 5,52. La fonction puissance est obtenue en calculant $P_\mu(\overline{X} >$5,52) avec $\overline{X} \rightsquigarrow \mathcal{N}(\mu\,, \frac{1}{10})$ pour $\mu \in]5, +\infty[$, c'est-à-dire en calculant $P(Z > \sqrt{10}(5{,}52-\mu)\,)$ où $Z \rightsquigarrow \mathcal{N}(0\,;1)$, soit $1-\Phi(\sqrt{10}(5{,}52-\mu))$ où Φ est la fonction de répartition de cette loi. Cette valeur croît avec μ comme le montre la figure 9.1. A gauche de $\mu = 5$ il s'agit du risque de première espèce. Cette croissance est caractéristique des familles à rapport de vraisemblance monotone. ■

Exemple 9.6 Soit la famille de lois exponentielles $\{f(x;\lambda) = \lambda e^{-\lambda x}; x \geq 0, \lambda > 0\}$. On a $c(\lambda) = -\lambda$ et $d(x) = x$. Pour $H_0 : \lambda \geq \lambda_0$ vs. $H_1 : \lambda < \lambda_0$ le test UPP repose sur $\sum_{i=1}^n x_i$ ou sur $\overline{x}$. Pour trouver le sens de l'inégalité on rappelle que la moyenne de la loi est $1/\lambda$. Ainsi une forte valeur observée pour $\overline{x}$ doit refléter une forte valeur de $1/\lambda$, soit une faible valeur de λ, ce qui est du côté de H_1. La région de rejet est donc de la forme $\overline{x} > k$.

La valeur de k doit être telle que $P_{\lambda_0}(\overline{X} > k) = \alpha$ pour obtenir le niveau α. Or, si $\lambda = \lambda_0$, $\sum_{i=1}^n X_i$ suit une loi $\Gamma(n, \lambda_0)$ (voir section 4.2.3). Comme $P_{\lambda_0}(\overline{X} > k) = P_{\lambda_0}(\sum_{i=1}^n X_i > nk)$, nk est le quantile d'ordre $1-\alpha$ sur la loi $\Gamma(n, \lambda_0)$, d'où l'on déduit k. ■

Pour la loi de Bernoulli $\mathcal{B}(p)$ qui régit notamment le test sur une proportion dans une population, on a $c(p) = \ln \frac{p}{1-p}$ qui est une fonction croissante de p et

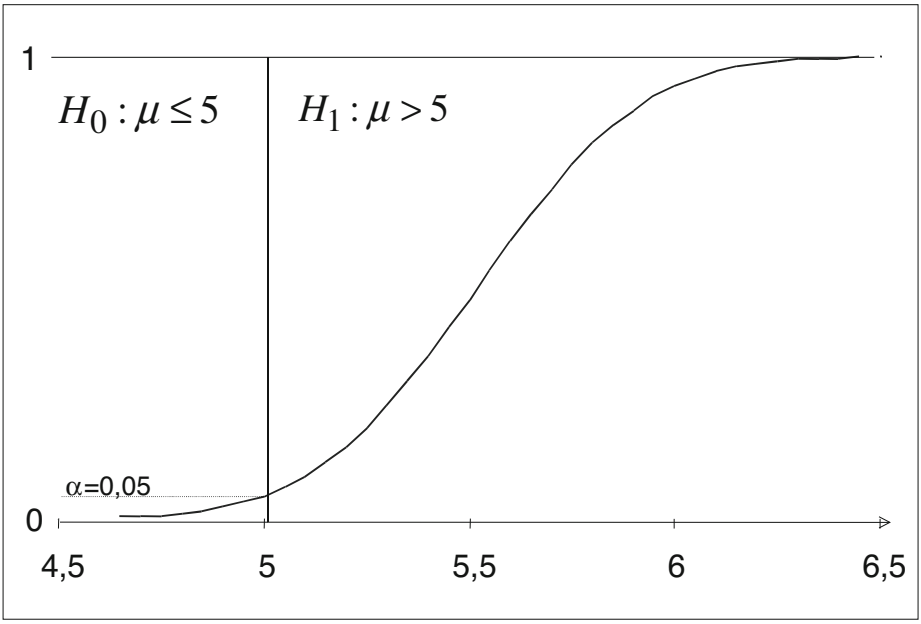


Figure 9.1 - Fonction puissance pour $H_0 : \mu = 5$ sur une loi $\mathcal{N}(\mu\,, 1)$.

$d(x) = x$ (voir exemple 9.4). Le test UPP de $H_0 : p \leq p_0$ vs. $H_1 : p > p_0$ repose donc sur $\sum_{i=1}^{n} x_i$ qui est le nombre de «succès» observé. On rejette H_0 si ce nombre est trop élevé (car cela fait pencher vers H_1) et la valeur critique k se lit sur la loi $\mathcal{BN}(n\,,\, p_0)$ comme dans l'exemple 9.4.

Pour la loi de Poisson $\mathcal{P}(\lambda)$ on a $c(\lambda) = \ln \lambda$ qui est une fonction croissante de λ et $d(x) = x$ (voir exemple 6.2). Pour déterminer la valeur critique k on utilise le fait que $\sum_{i=1}^{n} X_i$ suit une loi $\mathcal{P}(n\lambda)$. L'examen de la loi de Pareto est proposé dans les exercices.

L'existence d'un test UPP n'est pas réservée à la classe exponentielle. Pour la famille des lois uniformes $\mathcal{U}[0, \theta]$ on a vu que le maximum $X_{(n)}$ est statistique exhaustive minimale. On peut montrer qu'elle est à RV monotone et il existe donc un test UPP de la forme $X_{(n)} < k$ ou $X_{(n)} > k$ (voir exercices).

Remarque 9.1 Choix de H_0

Face à une situation pratique le choix du sens de H_0 ($\theta \leq \theta_0$ ou $\theta \geq \theta_0$) n'est pas toujours évident. Il devra se faire en considérant les deux erreurs possibles. On doit faire en sorte que celle qui est jugée la plus grave soit une erreur de première espèce : affirmer que H_0 est fausse (donc que l'on choisit H_1) alors

qu'elle est vraie. En d'autres termes l'affirmation la plus sensible doit correspondre à H_1. Dans l'illustration de l'exemple 9.5 on peut commettre une erreur soit en déclarant que la teneur n'est pas respectée alors qu'elle l'est pourtant, soit en déclarant qu'elle est respectée alors qu'elle ne l'est pas. La première erreur est beaucoup plus préjudiciable pour la personne effectuant le test (on ne veut pas accuser à tort). Aussi H_1 doit-elle exprimer le fait que la teneur n'est pas respectée, soit $H_1 : \mu > 5$ d'où $H_0 : \mu \leq 5$. Supposons qu'un médicament soit considéré efficace si un paramètre θ dépasse un seuil θ_0. On peut déclarer le médicament efficace alors qu'il ne l'est pas ou le déclarer inefficace alors qu'il est efficace. La première de ces erreurs est plus critique car elle aura pour conséquence de mettre sur le marché un médicament inutile, alors que pour la deuxième le médicament ne sera pas diffusé par mesure conservatoire. H_1 doit donc correspondre au fait que le médicament est efficace, soit $\theta > \theta_0$ d'où $H_0 : \theta \leq \theta_0$.

Notons que dans la théorie classique développée ici la notion de risque de première ou de deuxième espèce suppose que l'hypothèse nulle et, donc, l'hypothèse alternative aient été posées avant d'observer les données. En pratique il est fréquent que l'on décide du sens du rejet sur la base même des observations. Cette façon de faire est à rapprocher de l'usage de la P-valeur décrit en section 9.6 ou encore de ce qu'on appelle parfois un *test de significativité*.

9.4.3 Tests d'hypothèses bilatérales

Nous considérons deux situations du type bilatéral :

$$\begin{array}{lll} & H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0 \\ \text{ou} & H_0 : \theta_1 \leq \theta \leq \theta_2 \quad \text{vs.} \quad H_1 : \theta < \theta_1 \text{ ou } \theta > \theta_2 \end{array}$$

où θ est un paramètre de dimension 1 ($\Theta \subseteq \mathbb{R}$). La première situation est fréquente lorsque θ représente en fait un écart entre paramètres de deux populations, par exemple entre leurs moyennes (voir section 9.7.3). La seconde teste si le paramètre est situé dans un intervalle de tolérance acceptable. L'appellation de bilatéral se réfère au fait que l'alternative est située de part et d'autre de l'hypothèse nulle.

On ne peut s'attendre dans ces situations à obtenir un test UPP du fait qu'il faut faire face à des alternatives à la fois du type $\theta < \theta_0$ et du type $\theta > \theta_0$, par exemple pour le premier cas. Toutefois il pourra y avoir un test uniformément plus puissant dans la classe restreinte des tests sans biais, en bref **UPP-sans biais**. Ceci est notamment vrai pour les familles de lois de la classe exponentielle. D'une façon générale la région d'acceptation aura la forme $c_1 < t < c_2$, où $c_1 < c_2$, pour la réalisation d'une statistique de test T appropriée (soit $\sum_{i=1}^{n} d(X_i)$ pour la classe exponentielle). Ceci justifie l'appellation fréquente de *test bilatéral* puisqu'on est amené à un rejet sur les deux extrémités. Nous rencontrerons de telles situations dans la section sur les tests usuels.

Note 9.3 L'usage veut que l'on détermine les valeurs critiques c_1 et c_2 en répartissant $\alpha/2$ sur chaque extrémité. Ainsi, pour le cas $H_0 : \theta = \theta_0$, ces valeurs seront telles que $P_{\theta_0}(T < c_1) = P_{\theta_0}(T > c_2) = \alpha/2$. Mais cette règle ne conduit pas au test UPP-sans biais si la loi de T n'est pas symétrique (le test peut même ne plus être sans biais). Dans la classe exponentielle la répartition doit être telle que la dérivée par rapport à θ de $P_\theta(T < c_1) + P_\theta(T > c_2)$ s'annule en θ_0.

Pour le cas $H_0 : \theta_1 \leq \theta \leq \theta_2$ la condition est que le seuil α soit atteint à la fois en θ_1 et en θ_2. La résolution de tels problèmes n'est pas simple et, dès lors, la règle de répartition égale apparaît bien commode (une situation de test de ce type sera envisagée dans l'exemple 9.8).

Nous n'approfondirons pas la recherche de tests optimaux dans les cas bilatéraux et nous nous contenterons de présenter un test de portée générale, inspiré du rapport de vraisemblance simple, s'appliquant aux situations les plus complexes et, en particulier, aux hypothèses bilatérales ci-dessus.

9.5 Test du rapport de vraisemblance généralisé

Considérons maintenant les hypothèses paramétriques les plus générales.

Définition 9.11 *Soit la famille paramétrique* $\{f(x;\theta), \theta \in \Theta\}$, *où* $\Theta \subseteq \mathbb{R}^k$, *et les hypothèses* $H_0 : \theta \in \Theta_0$ *vs.* $H_1 : \theta \in \Theta_1$ *où* $\Theta_1 = \Theta - \Theta_0$ *est le complémentaire de* Θ_0 *par rapport à* Θ. *On appelle* ***rapport de vraisemblance généralisé*** *(RVG), la fonction* $\lambda(x_1, x_2, ..., x_n)$ *telle que :*

$$\lambda(x_1, x_2, ..., x_n) = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta; x_1, x_2, ..., x_n)}{\sup\limits_{\theta \in \Theta} L(\theta; x_1, x_2, ..., x_n)}$$

et ***test du RVG****, le test défini par une région de rejet de la forme :*

$$\lambda(x_1, x_2, ..., x_n) < k \leq 1.$$

Il est évident que $\lambda(x_1, x_2, ..., x_n)$ est inférieur ou égal à 1 pour toute réalisation $(x_1, x_2, ..., x_n)$. De plus, s'il existe une estimation du maximum de vraisemblance $\widehat{\theta}^{MV}$, et c'est pour ainsi dire toujours le cas (voir section 6.7.1), alors le dénominateur est la valeur de la fonction de vraisemblance en $\widehat{\theta}^{MV}$, soit : $L(\widehat{\theta}^{MV}; x_1, x_2, ..., x_n)$.

Le RVG relève de la même rationalité que le RV simple. Si, pour une réalisation donnée, la vraisemblance atteint un maximum dans H_0 qui reste bien inférieur à son maximum absolu dans tout l'espace paramétrique Θ, alors il y a lieu de douter de cette hypothèse.

Vérifions que dans le cas d'hypothèses nulle et alternative simples, le test du RVG est équivalent au test du RV simple. Le dénominateur du RV est $L(\theta_1; x_1, x_2, ..., x_n)$ alors que celui du RVG est la valeur maximale entre $L(\theta_0; x_1, x_2, ..., x_n)$ et $L(\theta_1; x_1, x_2, ..., x_n)$. Pour les réalisations $(x_1, x_2, ..., x_n)$ telles que le RV simple est strictement inférieur à 1, cette valeur maximale est donc atteinte pour θ_1 et le RV simple est égal au RVG et, réciproquement, si le RVG est strictement inférieur à 1 alors il en va de même pour le RV. Les régions de test $\lambda(x_1, x_2, ..., x_n) < k$ et $L(\theta_0; x_1, x_2, ..., x_n)/L(\theta_1; x_1, x_2, ..., x_n) < k$ sont donc identiques. Toutefois si le RV simple est supérieur à 1 alors la valeur maximale est atteinte pour θ_0, le RVG reste égal à 1 et il n'y a donc pas équivalence sur un plan strictement mathématique. Cependant il est clair qu'en règle générale le test du RV simple n'a de sens que s'il se fonde sur une valeur de k inférieure à 1, ceci pour garantir un risque de première espèce faible (voir le commentaire à la suite de la définition 9.6), et on peut considérer qu'il y a équivalence du point de vue pratique.

Le test du RVG n'a pas de propriétés d'optimalité notables mais on constate dans des situations usuelles qu'il donne le test UPP-sans biais (voir l'exemple 9.7). Cependant il possède des propriétés asymptotiques intéressantes, notamment sa convergence moyennant des conditions de régularité analogues à celles de l'estimateur du maximum de vraisemblance.

Le problème est toutefois de connaître la loi de la **statistique du RVG** $\lambda(X_1, X_2, \cdots, X_n)$ pour toute valeur de θ dans Θ_0 afin de définir la valeur de k permettant de garantir le niveau α choisi. En effet cette valeur doit être telle que :

$$\sup_{\theta \in \Theta_0} P_\theta(\lambda(X_1, X_2, \cdots, X_n) < k) = \alpha\,.$$

Il arrive, comme dans l'exemple qui suit, que la région de rejet se ramène à une forme simple. Mais cela reste l'exception et nous verrons plus loin qu'on dispose, sinon, d'une approximation asymptotique très utile.

Exemple 9.7 Soit la famille de loi $\mathcal{N}(\mu, \sigma^2)$ avec (μ, σ^2) inconnu et l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$. Comme μ est inconnu H_0 et H_1 sont toutes deux multiples. Pour (μ, σ^2) quelconque la fonction de vraisemblance est :

$$\begin{aligned} L(\mu, \sigma^2; x_1, x_2, \cdots, x_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\}\,. \end{aligned}$$

On a vu (exemple 6.20) que l'estimation du MV pour le paramètre (μ, σ^2) inconnu est $(\overline{x}, \widetilde{s}^2)$ où $\widetilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Le dénominateur du RVG est donc obtenu en remplaçant respectivement μ et σ^2 par $\overline{x}$ et $\widetilde{s}^2$. Pour le numérateur il s'agit de maximiser uniquement sur μ car σ^2 est fixé, égal à σ_0^2. Cela revient

à minimiser $\sum_{i=1}^n (x_i-\mu)^2$ ce qui s'obtient pour $\mu=\overline{x}$. Finalement le RVG est égal à :

$$\lambda(x_1,x_2,...,x_n)=\frac{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\overline{x})^2\}}{\left(\frac{1}{\sqrt{2\pi\widetilde{s}^2}}\right)^n \exp\{-\frac{1}{2\widetilde{s}^2}\sum_{i=1}^n (x_i-\overline{x})^2\}}$$
$$=\left(\frac{\widetilde{s}^2}{\sigma_0^2}\right)^{\frac{n}{2}} \frac{\exp\{-\frac{n}{2}\frac{\widetilde{s}^2}{\sigma_0^2}\}}{\exp\{-\frac{n}{2}\}}=\left(\frac{\widetilde{s}^2}{\sigma_0^2}\exp\{1-\frac{\widetilde{s}^2}{\sigma_0^2}\}\right)^{\frac{n}{2}}.$$

Considérons la fonction $g(u)=u\exp\{1-u\}$. Elle est nulle pour $u=0$ et tend vers 0 quand $u\to+\infty$. Son sens de variation est le même que celui de $\ln g(u)=\ln u+1-u$ dont la dérivée est $\frac{1}{u}-1$, laquelle est positive pour $u<1$, s'annule en $u=1$ et est négative pour $u>1$. La fonction $g(u)$ admet donc un unique maximum sur $[0,+\infty[$ en $u=1$. La région de rejet $\lambda(x_1,x_2,...,x_n)<k$ se traduit ainsi en $u_1<\frac{\widetilde{s}^2}{\sigma_0^2}<u_2$ où $u_1<1<u_2$ et $g(u_1)=g(u_2)$, soit encore, en multipliant par n :

$$nu_1<\frac{\sum_{i=1}^n (x_i-\overline{x})^2}{\sigma_0^2}<nu_2\,.$$

Or la statistique $\sum_{i=1}^n (X_i-\overline{X})^2/\sigma_0^2$ suit une loi $\chi^2(n-1)$ (voir théorème 5.1) ce qui permet de trouver des quantiles appropriés $\chi^{2\,(n-1)}_{\alpha_1}$ et $\chi^{2\,(n-1)}_{1-\alpha_2}$. Ceux-ci, pour le niveau α choisi, doivent d'une part vérifier $\alpha_1+\alpha_2=\alpha$ et d'autre part $g(\frac{1}{n}\chi^{2\,(n-1)}_{\alpha_1})=g(\frac{1}{n}\chi^{2\,(n-1)}_{1-\alpha_2})$. Ce test est de la même forme que le test classique que nous verrons en section 9.7.2 et dont on peut montrer qu'il est UPP-sans biais pour un choix particulier des quantiles ci-dessus. ■

Exemple 9.8 Nous donnons maintenant un exemple dans la situation :

$$H_0:\theta_1\le\theta\le\theta_2 \qquad \text{vs.} \qquad H_1:\theta<\theta_1 \text{ ou } \theta>\theta_2$$

qui, bien qu'assez peu envisagée par les praticiens, est souvent plus réaliste que celle où $H_0:\theta=\theta_0$. En effet tester une valeur ponctuelle n'a pas grand sens tant il est vrai qu'on peut être certain qu'elle ne peut correspondre de façon exacte à la vérité. D'ailleurs, pour un test convergent, on sera amené à coup quasi sûr à rejeter cette hypothèse avec de grands échantillons. Dans la mesure du possible il est préférable de considérer une marge de tolérance $[\theta_1,\theta_2]$ pour le paramètre inconnu.

Nous prenons le cas d'une loi mère $\mathcal{N}(\mu,\sigma^2)$ où le paramètre (μ,σ^2) de dimension 2 est inconnu et testons donc :

$$H_0:\mu_1\le\mu\le\mu_2 \quad \text{vs.} \quad H_1:\mu<\mu_1 \text{ ou } \mu>\mu_2.$$

Le dénominateur du RVG vaut $\left(\sqrt{2\pi\widetilde{s}^2}\right)^{-n}\exp\{-\frac{n}{2}\}$ comme dans l'exemple 9.7. Pour trouver le numérateur il faut maximiser :

$$\left(\sqrt{2\pi\sigma^2}\right)^{-n}\exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\}$$

pour $\mu\in[\mu_1,\mu_2]$ et $\sigma^2>0$ quelconque. En passant au logarithme cela revient à minimiser :

$$\ln\sigma^2+\frac{1}{\sigma^2}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

Pour a fixé, $\ln\sigma^2+\frac{1}{\sigma^2}a$ est minimal quand σ^2 est égal à a et il faut donc minimiser la fonction :

$$\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}x_i^2+n\mu(\mu-2\overline{x})$$

ou encore $h(\mu)=\mu(\mu-2\overline{x})$.

Pour les réalisations telles que $\overline{x}$ soit dans $[\mu_1,\mu_2]$ le minimum est atteint pour $\mu=\overline{x}$. Mais dans ce cas la solution est identique à celle du dénominateur et le RVG vaut 1. On est donc nécessairement dans la zone d'acceptation (l'inégalité pour la zone de rejet donnée dans la définition 9.11 étant stricte).

Si $\overline{x}<\mu_1$ le minimum cherché pour $\mu\in[\mu_1,\mu_2]$ est obtenu pour $\mu=\mu_1$ car la fonction $h(\mu)$ est croissante pour $\mu>\overline{x}$. Le numérateur du RVG vaut alors $\left(\sqrt{2\pi s_1^2}\right)^{-n}\exp\{-\frac{n}{2}\}$ où $s_1^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_1)^2$ et le RVG vaut $(s_1^2/\widetilde{s}^2)^{-n/2}$. L'inégalité $\lambda(x_1,x_2,...,x_n)<k$ se traduit donc en $s_1^2/\widetilde{s}^2>k'$ avec $k'>1$. Or :

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_1)^2&=\frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\overline{x})+(\overline{x}-\mu_1)\right]^2\\&=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2+\frac{2}{n}(\overline{x}-\mu_1)\sum_{i=1}^{n}(x_i-\overline{x})+\frac{1}{n}\sum_{i=1}^{n}(\overline{x}-\mu_1)^2\\&=\widetilde{s}^2+(\overline{x}-\mu_1)^2.\end{aligned}$$

Donc $s_1^2/\widetilde{s}^2=1+(\overline{x}-\mu_1)^2/\widetilde{s}^2$ et $s_1^2/\widetilde{s}^2>k'$ équivaut à $(\overline{x}-\mu_1)^2/\widetilde{s}^2>k''$ où $k''>0$, soit encore, après multiplication par $n-1$ et en prenant la racine carrée, $|\overline{x}-\mu_1|/(s/\sqrt{n})>k_1$ avec $k_1>0$. Comme $\overline{x}<\mu_1$ on a finalement une région de rejet de la forme $(\overline{x}-\mu_1)/(s/\sqrt{n})<-k_1$.

De la même façon on obtient, pour les réalisations telles que $\overline{x}>\mu_2$, la région de rejet $(\overline{x}-\mu_2)/(s/\sqrt{n})>k_2$. La difficulté vient alors du fait qu'il faut trouver les constantes positives k_1 et k_2 telles que, pour tout $\mu\in[\mu_1,\mu_2]$:

$$P_\mu\left(\frac{\overline{X}-\mu_1}{S/\sqrt{n}}<-k_1\right)+P_\mu\left(\frac{\overline{X}-\mu_2}{S/\sqrt{n}}>k_2\right)\leq\alpha$$

sachant que $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ suit une loi de Student à n-1 degrés de liberté. Même en admettant que le risque maximal est atteint en μ_1 ou en μ_2, la solution n'est pas simple. En effet si, par exemple, $\mu = \mu_1$ la deuxième probabilité concerne une v.a. qui ne suit pas une loi de Student classique du fait que μ_2 n'est pas la moyenne de $\overline{X}$, mais une *loi de Student non centrale* de paramètre de non-centralité $(\mu_1 - \mu_2)/\sigma$ dont les tables sont peu répandues (sans compter que σ est inconnu). On peut mettre en évidence une **procédure conservatrice** en réécrivant la somme des probabilités ci-dessus, dans ce même cas où $\mu = \mu_1$, selon :

$$P_{\mu_1}\left(\frac{\overline{X}-\mu_1}{S/\sqrt{n}} < -k_1\right) + P_{\mu_1}\left(\frac{\overline{X}-\mu_1}{S/\sqrt{n}} > k_2 + \frac{\mu_2-\mu_1}{S/\sqrt{n}}\right).$$

En prenant $k_1 = k_2 = t^{(n-1)}_{1-\alpha/2}$, la première probabilité vaut bien $\alpha/2$ et la seconde est certainement inférieure à $\alpha/2$ puisque $\mu_2 - \mu_1$ est positif. Le raisonnement est identique lorsque $\mu = \mu_2$. Notons que la procédure est d'autant plus conservatrice que les deux moyennes s'éloignent.

La situation $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ peut être considérée comme un cas particulier de la situation ci-dessus, avec $\mu_1 = \mu_2 = \mu_0$. On a alors une procédure exacte en choisissant pour $-k_1$ un quantile $t^{(n-1)}_{\alpha_1}$ de la loi $t(n-1)$ et pour k_2 un quantile $t^{(n-1)}_{1-\alpha_2}$ de façon que $\alpha_1+\alpha_2=\alpha$. En raison de la symétrie de la loi de Student il semble naturel d'opter pour les quantiles symétriques $t^{(n-1)}_{\alpha/2} = -t^{(n-1)}_{1-\alpha/2}$ et $t^{(n-1)}_{1-\alpha/2}$. Ceci correspond au test classique présenté en section 9.7.1 dont on montre qu'il est UPP-sans biais.

En exercices est proposé le cas plus simple où σ^2 est connu. ■

Paramètre de nuisance

Supposons que la loi mère appartienne à une famille à paramètre de dimension 2, noté pour la circonstance (θ, ρ) où θ et ρ sont ses deux composantes. Si l'hypothèse nulle ne spécifie que la composante θ, la composante ρ est appelée *paramètre de nuisance.* C'est le cas de σ^2 dans l'exemple ci-dessus ou celui de μ dans l'exemple 9.7. Pour les tests usuels, la présence d'un paramètre de nuisance fera qu'il n'y aura généralement pas de test UPP, mais il peut exister un test UPP-sans biais. Ceci est vrai pour l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ de l'exemple 9.7 ou pour $H_0 : \mu = \mu_0$ de l'exemple qui précède. Pour une hypothèse nulle de la forme $H_0 : \theta_1 \leq \theta \leq \theta_2$ on montre qu'il existe un test UPP-sans biais pour une famille de la classe exponentielle si sa densité (ou fonction de probabilité) peut s'écrire :

$$f(x;\theta,\rho) = a(\theta,\rho)\, b(x)\, \exp\{c_1(\theta)d_1(x) + c_2(\rho)d_2(x)\}.$$

On notera que ceci n'est pas vérifié par la loi de Gauss qui ne sépare pas ainsi μ et σ^2 dans la partie exponentielle. De fait, il n'existe pas de test UPP-sans

biais pour $H_0 : \mu_1 \leq \mu \leq \mu_2$ considérée ci-dessus. Ces résultats se généralisent à plusieurs paramètres de nuisance.

Nous donnons maintenant un résultat asymptotique très précieux qui permettra de déterminer une région de rejet approchée dans le cas de grands échantillons.

Théorème 9.2 *Soit la famille paramétrique* $\{f(x;\theta), \theta \in \Theta\}$, *où* $\Theta \subseteq \mathbb{R}^k$, *et l'hypothèse* H_0 *spécifiant les valeurs de* r *composantes de* θ $(1 \leq r \leq k)$. *Supposons que soient remplies les conditions de régularité garantissant que l'estimateur du maximum de vraisemblance soit BAN (voir proposition 6.11). Alors, sous* H_0, *la statistique du RVG* $\Lambda_n = \lambda(X_1, X_2, \cdots, X_n)$ *est telle que :*

$$-2\ln\Lambda_n \xrightarrow{\mathcal{L}} \chi^2(r).$$

Donnons une esquisse de démonstration dans le cas simple où le paramètre θ inconnu est de dimension $k = 1$ et est donc parfaitement spécifié par H_0 : $\theta = \theta_0$. On verra ainsi que ce résultat est lié aux propriétés asymptotiques de l'estimateur du MV. Pour une quelconque réalisation $(x_1, x_2, \cdots, x_n)$ notée en bref $\mathbf{x}_n$, développons en série de Taylor la log-vraisemblance de $\mathbf{x}_n$ en θ_0 autour de l'estimation du maximum de vraisemblance $\widehat{\theta}_n$:

$$\ln L(\theta_0;\mathbf{x}_n) = \ln L(\widehat{\theta}_n;\mathbf{x}_n) + (\widehat{\theta}_n - \theta_0)\frac{\partial}{\partial\theta}\ln L(\widehat{\theta}_n;\mathbf{x}_n) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^2\frac{\partial^2}{\partial\theta^2}\ln L(\widetilde{\theta}_n;\mathbf{x}_n)$$

où $\widetilde{\theta}_n$ est une valeur comprise entre θ_0 et $\widehat{\theta}_n$. Comme, par définition de l'estimation du MV, $\frac{\partial}{\partial\theta}\ln L(\widehat{\theta}_n;\mathbf{x}_n) = 0$, on a, pour le RVG :

$$-2\ln\lambda_n = -2[\ln L(\theta_0;\mathbf{x}_n) - \ln L(\widehat{\theta}_n;\mathbf{x}_n)],$$

d'où :

$$-2\ln\lambda_n = -(\widehat{\theta}_n - \theta_0)^2\frac{\partial^2}{\partial\theta^2}\ln L(\widetilde{\theta}_n;\mathbf{x}_n).$$

En passant aux v.a. (avec $\mathbf{X}_n = (X_1, X_2, \cdots, X_n)$), sous l'hypothèse nulle, $\widehat{\theta}_n$ converge en probabilité vers θ_0 et il en va donc de même pour $\widetilde{\theta}_n$. On peut alors montrer, d'une part par continuité de $\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)$ par rapport à θ (condition de régularité), d'autre part par la loi des grands nombres, que :

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ln L(\widetilde{\theta}_n;\mathbf{X}_n) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ln f(X_i;\widetilde{\theta}_n)$$

converge en probabilité et donc en loi vers $E_{\theta_0}\left[-\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta_0)\right] = I(\theta_0)$. Ainsi $-2\ln\Lambda_n$ a la même convergence en loi que la suite de v.a. $nI(\theta_0)(\hat{\theta}_n - \theta_0)^2$. Or,

selon la proposition 6.11, $\sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta)$ converge en loi vers la loi $\mathcal{N}(0\,;1)$ d'où il découle que $nI(\theta_0)(\hat{\theta}_n - \theta_0)^2$ converge en loi vers la loi $\chi^2(1)$. □

Comme la région de rejet $\lambda < k$ est équivalente à $-2\ln\lambda > k'$ on rejettera à un niveau approximatif α si :

$$-2\ln\lambda > \chi^{2\ (r)}_{1-\alpha}.$$

Ce résultat dont la validité s'étend au-delà de l'échantillonnage aléatoire simple autorise un test approché dans des situations complexes. C'est pourquoi **on trouve le test du RVG de façon omniprésente dans les logiciels.** Nous montrerons plus loin qu'il est à la base des tests classiques dits du khi-deux portant sur des fréquences (voir sections 10.1 à 10.4).

Note 9.4 Dans les situations de test que nous avons décrites, le cas le plus fréquent est celui d'une hypothèse nulle ne spécifiant qu'une seule dimension du paramètre, la statistique du RVG suivant alors une loi $\chi^2(1)$. Quand le RVG est un test usuel avec une statistique dont la loi est connue sous H_0 on peut apprécier la validité de l'approximation asymptotique (voir exercice 9.8). Notons que la région de rejet au niveau 0,05 est $-2\ln\lambda > \chi^{2\ (1)}_{0,95} = 3{,}84$ ce qui signifie un rapport de vraisemblance inférieur à 0,147.

Note 9.5 Lorsque l'on comparera plusieurs populations on posera l'égalité de certains paramètres sans vraiment les spécifier tous. On verra par exemple en section 9.7.3 l'hypothèse nulle d'égalité des moyennes de deux lois gaussiennes $H_0 : \mu_1 = \mu_2$. En fait pour le théorème ci-dessus on peut considérer que l'on spécifie un paramètre. Pour le voir il suffit de poser $\mu_2 = \mu_1 + \theta$ et H_0 s'écrit alors $\theta = 0$.

9.6 Remarques diverses

La plupart des tests disponibles ont été développés à partir d'idées intuitives en imaginant une statistique dont le comportement est très différencié sous H_0 et sous H_1, et dont la loi exacte ou approchée est accessible sous H_0. Evidemment de tels tests seront rarement UPP. Pour une situation de test donnée il existe généralement, comme on peut le voir dans les logiciels, plusieurs propositions de test. Il est difficile de savoir quel test est préférable car le calcul formel de la puissance ne peut être conduit, soit parce que la loi de la statistique est trop complexe pour l'alternative, soit en raison de la multiplicité de formes de cette alternative (notamment pour les tests de type non paramétrique comme on en verra au chapitre 10). Suite à des études examinant des hypothèses alternatives particulières, formellement ou par simulation, on dispose parfois d'éléments permettant d'effectuer le meilleur choix compte tenu des formes les plus plausibles de l'alternative dans la situation considérée.

Comme pour les intervalles de confiance les résultats asymptotiques concernant l'estimateur du MV peuvent être utilisés pour construire un test approximatif. En effet $\sqrt{nI(\theta)}\,(\hat{\theta}_n^{MV} - \theta)$ suit approximativement une loi $\mathcal{N}(0\,;1)$. Si l'hypothèse nulle spécifie parfaitement $\theta = \theta_0$ alors $\sqrt{nI(\theta_0)}\,(\hat{\theta}_n^{MV} - \theta_0)$ est une statistique de loi approximativement connue, ce qui permet de définir une région de rejet. Ceci sera utilisé dans la prochaine section pour le cas d'une loi mère de Bernoulli avec de grands échantillons.

La théorie classique des tests peut être généralisée dans le cadre de la théorie de la décision. Celle-ci stipule une *fonction de perte* qui définit un coût pour l'erreur de première et pour celle de deuxième espèce, puis la notion de risque comme espérance mathématique de cette fonction de perte. Le risque étant une fonction de θ l'objectif est alors de rechercher le test qui minimise le risque, si possible uniformément en θ. Un tel test n'existant généralement pas on se satisfera, par exemple, d'un *test minimax* pour lequel le risque maximum sur Θ reste inférieur ou égal à celui de tout autre test. On peut conjuguer cette approche avec une **approche bayésienne** en considérant une loi *a priori* sur Θ et en minimisant le risque *a posteriori* qui ne dépend plus du ou des paramètres inconnus. Notons que la théorie classique repose implicitement sur une fonction de perte attribuant le coût 1 à une erreur de première comme de deuxième espèce.

L'usage courant de la P-valeur

La décision d'accepter ou de refuser une hypothèse est sujette au choix du risque de première espèce α. Afin d'éviter ce choix on peut recourir, et c'est ce que font les logiciels, à la notion de P-valeur pour simplement rendre compte du résultat d'un test. **La P-valeur est la probabilité que, sous H_0, la statistique de test prenne une valeur au moins aussi extrême que celle qui a été observée**. La notion de position extrême se définit en relation avec la définition du seuil du test. Si la région de rejet est unilatérale du type $t > c$, alors pour une valeur t_0 observée après expérience la P-valeur est $P(T > t_0\,|H_0)$ si H_0 est simple ou bien le maximum de $P_\theta(T > t_0)$ sur Θ_0 si elle est multiple. Si la région de rejet est bilatérale, par exemple $\{t\,|\,t < c_1$ ou $t > c_2\}$ alors la P-valeur est définie par $2P(T < t_0\,|H_0)$ si t_0 est plus petit que la médiane de la loi de T sous H_0 ou $2P(T > t_0\,|H_0)$ s'il est plus grand, ceci afin de tenir compte du rejet sur les deux extrémités.

Reprenons la situation de l'exemple 9.5 avec $H_0 : \mu \leq 5$ vs. $H_1 : \mu > 5$ pour une loi mère $\mathcal{N}(\mu\,,\,1)$. On a vu que sur la base d'un échantillon de taille 10 on est amené à rejeter H_0 si $\overline{x} > 5 + 1{,}645(1/\sqrt{10})$ au niveau 0,05 ou, plus généralement, $\overline{x} > 5 + z_{1-\alpha}(1/\sqrt{10})$ au niveau α où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Gauss centrée-réduite. Supposons que l'échantillon observé ait une moyenne égale à 6. Comme le risque de première espèce maximal est atteint pour $\mu = 5$ la P-valeur est égale à $P(\overline{X} > 6)$ pour $\overline{X} \rightsquigarrow \mathcal{N}(5\,;\,1/10)$ soit $P(Z > \sqrt{10}(6-5)\,) = P(Z > 3,16)$ pour $Z \rightsquigarrow \mathcal{N}(0\,;\,1)$, laquelle est égale à

0,008 ce qui indique directement que la valeur observée est au-delà de la valeur critique au niveau 0,05 et même au niveau 0,01. Si le test avait été bilatéral avec $H_0 : \mu = 5$ vs. $H_1 : \mu \neq 5$, la P-valeur correspondant à la même observation 6 aurait été prise égale à 0,016 impliquant un rejet au niveau 0,05 mais pas au niveau 0,01.

Avec cette définition, d'une façon générale, la P-valeur permet de déterminer si l'on rejette à un niveau α donné (à condition toutefois que, dans le cas bilatéral, la zone de rejet soit partagée en risque $\alpha/2$ équitablement sur chaque extrémité, ce qui est l'usage courant). Si la P-valeur est inférieure à α on rejette H_0 sinon on l'accepte. Comme autre façon de voir les choses on peut dire que plus la P-valeur est faible plus l'hypothèse nulle est suspecte. Ainsi l'indication des P-valeurs dans les logiciels a rendu obsolète l'usage des tables pour le praticien.

9.7 Les tests paramétriques usuels

Certains de ces tests ont, en fait, déjà été vus dans la théorie générale et nous les indiquerons plus brièvement. La construction des tests usuels découle de la présence d'une statistique de loi connue sous H_0 et souvent sous H_1, s'imposant de façon assez naturelle, indépendamment de toute considération d'optimalité. Dans la plupart des cas il se trouve que le test ainsi construit est UPP-sans biais ce que nous mentionnerons au passage.

Il y a un parallélisme étroit entre les sections 7.4.1 à 7.4.6 pour la construction des intervalles de confiance usuels et les sections qui vont suivre. En effet le point de départ est identique. Pour un IC on met en évidence une statistique T telle qu'il existe une fonction $g(T, \theta)$ dont la loi est indépendante de θ, ce qui permet un encadrement à un niveau de probabilité souhaité. Pour peu que cette fonction puisse pivoter (voir définition 7.3) on en déduit un encadrement de θ. Pour un test les choses sont plus simples car on spécifie $\theta = \theta_0$ sous H_0 et $g(T, \theta_0)$ devient une statistique de loi connue. Soit $u_{\alpha/2}$ et $u_{1-\alpha/2}$ les quantiles d'ordres respectifs $\alpha/2$ et $1-\alpha/2$ pour cette loi, on peut donner comme région d'acceptation $A = \{t \,|\, g(t, \theta_0) \in [u_{\alpha/2}, u_{1-\alpha/2}]\}$. Il n'y a donc pas nécessité à pivoter pour la fonction $g(t, \theta)$ mais, en revanche, il faut souhaiter qu'elle soit monotone vis-à-vis de t pour que A soit un intervalle pour t. Le calcul de la puissance, par exemple en $\theta = \theta_1 : P_{\theta_1}(g(T, \theta_0) \notin [u_{\alpha/2}, u_{1-\alpha/2}])$, pourra toutefois poser des difficultés dans la mesure où la statistique est ici $g(T, \theta_0)$ et non pas $g(T, \theta_1)$: la loi de $g(T, \theta_0)$ «sous θ_1» n'aura pas souvent une forme simple. Quoi qu'il en soit tous les tests mis en évidence ci-après seront convergents.

En fin de chapitre nous reviendrons plus précisément sur le lien entre IC et tests et montrerons l'exploitation que l'on peut en faire. Nous proposons aussi dans la section des exercices quelques « exercices appliqués » permettant de voir des situations pratiques pour les tests usuels.

9.7.1 Tests sur la moyenne d'une loi $\mathcal{N}(\mu, \sigma^2)$

Cas où σ^2 est connu

Considérons tout d'abord ce cas d'école simple. La statistique T du cas général évoqué ci-dessus est la statistique exhaustive minimale $\overline{X}$ et la fonction $g(T, \theta)$ est $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ de loi connue : $\mathcal{N}(0\,;1)$. Notons que ce point de départ est celui de la section 7.4.1. Ainsi pour une hypothèse nulle $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ on a, sous H_0 :

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0\,;1).$$

Comme :

$$\begin{aligned} 1 - \alpha &= P_{\mu_0}\left(-z_{1-\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= P_{\mu_0}\left(\mu_0 - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

une région d'acceptation peut donc être définie, pour un test de risque α, par :

$$A = \{\overline{x} \mid \mu_0 - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \overline{x} < \mu_0 + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\}.$$

Ce test est UPP-sans biais, en vertu de la propriété mentionnée en section 9.4.3 pour la classe exponentielle. Ici la fonction puissance $h(\mu)$ peut être déterminée car la loi de $(\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ reste gaussienne quand μ est différent de μ_0. En effet :

$$\begin{aligned} h(\mu) &= 1 - P_\mu\left(-z_{1-\alpha/2} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= 1 - P_\mu\left(-z_{1-\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) \\ &= 1 - P_\mu\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) \\ &= 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\right) + \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2}\right) \end{aligned}$$

où Φ est la fonction de répartition de la loi de Gauss centrée-réduite. On montre que la fonction $h(\mu)$ s'accroît de part et d'autre de μ_0 à partir de la valeur α (voir une illustration dans l'exercice 9.3). Par ailleurs on vérifie aisément que $h(\mu) \to 1$ quand $n \to \infty$ aussi bien lorsque $\mu > \mu_0$ que lorsque $\mu < \mu_0$ ce qui démontre la convergence du test.

Pour des hypothèses unilatérales, par exemple $H_0 : \mu \leq \mu_0$ vs. $H_1 : \mu > \mu_0$, il est naturel de rejeter H_0 lorsque $\overline{x}$ est trop grand car cela reflète une moyenne

μ élevée. Pour déterminer la valeur critique on se place en $\mu = \mu_0$ qui est la valeur la plus défavorable pour H_0 (plus précisément, comme on l'a vu en section 9.4.2, le risque de première espèce est maximal en $\mu = \mu_0$). Comme :

$$\alpha = P_{\mu_0}\left(z_{1-\alpha} < \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right),$$

on a pour région de rejet $\overline{A} = \{\overline{x} \mid \overline{x} \geq \mu_0 + z_{1-\alpha}\,\sigma/\sqrt{n}\}$. Pour $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$ la région de rejet sera $\overline{A} = \{\overline{x} \mid \overline{x} < \mu_0 - z_{1-\alpha}\,\sigma/\sqrt{n}\}$.

Ces tests unilatéraux sont UPP comme vu en section 9.4.2 (voir exemple 9.5).

Cas où σ^2 est inconnu

Passons maintenant à ce cas général et plus réaliste. Ici une statistique exhaustive minimale est nécessairement de dimension 2 et l'on prendra $(\overline{X}, S^2)$. Mais on dispose d'une fonction à valeur dans $\mathbb{R}$ de loi connue quel que soit μ, à savoir $(\overline{X} - \mu)/(S/\sqrt{n})$ de loi de Student $t(n-1)$. Ainsi pour une hypothèse nulle $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ on a, sous H_0 :

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \rightsquigarrow t(n-1).$$

Comme :

$$1 - \alpha = P_{\mu_0}\left(-t^{(n-1)}_{1-\alpha/2} < \frac{\overline{X} - \mu_0}{S/\sqrt{n}} < t^{(n-1)}_{1-\alpha/2}\right),$$

une région de rejet peut être définie, pour un test de risque α, par :

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \notin [-t^{(n-1)}_{1-\alpha/2}\,,\, t^{(n-1)}_{1-\alpha/2}].$$

Pour des hypothèses bilatérales on aura, comme régions de rejet :

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} > t^{(n-1)}_{1-\alpha/2} \qquad \text{pour } H_0 : \mu \leq \mu_0\,,$$
$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} < -t^{(n-1)}_{1-\alpha/2} \qquad \text{pour } H_0 : \mu \geq \mu_0\,.$$

Le test bilatéral est UPP-sans biais de même que les tests unilatéraux (ces derniers ne sont qu'UPP-sans biais en raison de la présence du paramètre de nuisance σ^2).

Qu'en est-il de la détermination de la fonction puissance de ce *test de Student* ? Par transcription du cas σ^2 connu il nous faut calculer, pour μ «dans H_1» :

$$h(\mu) = P_{\mu}\left(-t^{(n-1)}_{1-\alpha/2} < \frac{\overline{X} - \mu_0}{S/\sqrt{n}} < t^{(n-1)}_{1-\alpha/2}\right).$$

Mais ici $(\overline{X} - \mu_0)/(S/\sqrt{n})$ ne suit plus une loi de Student car μ_0 n'est plus la moyenne de $\overline{X}$. En écrivant :

$$\frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{\overline{X} - \mu_0 + (\mu - \mu_0)}{S/\sqrt{n}}$$

on met en évidence le fait que cette statistique de test suit alors une loi de Student non centrale de paramètre de non centralité $(\mu - \mu_0)/\sigma$, déjà rencontrée dans l'exemple 9.8. Les tables des lois de Student non centrales sont volumineuses et peu répandues. Les logiciels pourraient faciliter la tâche mais rares sont ceux qui ont intégré ces calculs. Il existe également des abaques pour ce problème.

Pour les mêmes raisons que celles indiquées lors de l'étude de l'IC sur la moyenne (voir section 7.4.1), ce test est robuste vis-à-vis de l'hypothèse gaussienne. De fait les praticiens l'utilisent sans se soucier de cette hypothèse.

Pour ce qui concerne l'hypothèse :

$$H_0 : \mu_1 \leq \mu \leq \mu_2 \; vs. \; H_1 : \mu < \mu_1 \text{ ou } \mu > \mu_2$$

on a vu les difficultés de mise en oeuvre dans l'exemple 9.8. On a montré qu'en rejetant H_0 lorsque $(\overline{x} - \mu_1)/(s/\sqrt{n}) < -t_{1-\alpha/2}^{(n-1)}$ ou $(\overline{x} - \mu_2)/(s/\sqrt{n}) > t_{1-\alpha/2}^{(n-1)}$ on obtient un test conservateur de niveau α. Il n'existe pas de test UPP-sans biais dans ce cas.

9.7.2 Test sur la variance σ^2 d'une loi $\mathcal{N}(\mu, \sigma^2)$

Nous supposons que μ est inconnu et nous nous intéressons à une hypothèse sur σ^2, par exemple $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$. Ce test a déjà été traité dans l'exemple 9.7 via le rapport de vraisemblance généralisé. Nous allons retrouver le même test en développant l'approche directe analogue à celle de l'intervalle de confiance vue en section 7.4.2. Le point de départ est identique, à savoir le fait que :

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-1).$$

Sous H_0 on a donc :

$$P\left(\chi_{\alpha/2}^{2\,(n-1)} < \frac{(n-1)S^2}{\sigma_0^2} < \chi_{1-\alpha/2}^{2\,(n-1)}\right) = 1 - \alpha.$$

On dispose donc d'une statistique de test $(n-1)S^2/\sigma_0^2$ et d'une région d'acceptation $[\chi_{\alpha/2}^{2\,(n-1)}, \chi_{1-\alpha/2}^{2\,(n-1)}]$.

Ce test est UPP-sans biais, mais pour un choix précis de quantiles $\chi_{\alpha_1}^{2\,(n-1)}$ et $\chi_{1-\alpha_2}^{2\,(n-1)}$ tels que $\alpha_1 + \alpha_2 = \alpha$ (on montre que la condition à remplir, conformément à la note 9.3, est que la probabilité associée à l'intervalle

délimité par ces deux valeurs sur la loi $\chi^2(n+1)$ soit égale à 1-α). Ce choix est compliqué et l'on s'en tient généralement au choix ci-dessus qui est assez proche.

Pour l'hypothèse unilatérale $H_0 : \sigma^2 \leq \sigma_0^2$ la région de rejet doit être intuitivement $(n-1)S^2/\sigma_0^2 > \chi^{2\,(n-1)}_{1-\alpha}$ car une valeur élevée de S^2 rend H_0 suspecte. Pour $H_0 : \sigma^2 \geq \sigma_0^2$ on rejette si $(n-1)S^2/\sigma_0^2 < \chi^{2\,(n-1)}_{\alpha}$. Ces tests sont également UPP-sans biais.

Rappelons que le résultat sur la distribution d'échantillonnage de $(n-1)S^2/\sigma^2$ est peu robuste vis-à-vis de l'hypothèse gaussienne et, par conséquent, ces tests ne sont valables que si la loi mère est proche d'une loi de Gauss.

A titre de curiosité, si la moyenne μ était connue on utiliserait le fait que

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \rightsquigarrow \chi^2(n),$$

les développements étant analogues aux précédents.

9.7.3 Tests de comparaison des moyennes de deux lois de Gauss

On est en présence de **deux échantillons indépendants**, l'un de taille n_1, de moyenne $\overline{X}_1$ et variance S_1^2, issu d'une loi $\mathcal{N}(\mu_1, \sigma_1^2)$, l'autre de taille n_2, de moyenne $\overline{X}_2$ et variance S_2^2 issu d'une loi $\mathcal{N}(\mu_2, \sigma_2^2)$. En général les deux moyennes des lois et les deux variances sont inconnues. On souhaite comparer les deux moyennes μ_1 et μ_2 sur la base des échantillons. Essentiellement, les questions qui se posent sont de savoir si l'on peut décider à un niveau de risque α donné si elles sont différentes (cas bilatéral) ou si l'une est supérieure à l'autre (cas unilatéral). Ce type de situation, bilatérale ou unilatérale, est très fréquent car on est souvent amené à comparer deux populations réelles ou virtuelles suivant leurs moyennes. Par exemple dans l'expérimentation clinique on veut démontrer l'efficacité d'un traitement en comparant un échantillon traité et un échantillon témoin. Prenant soin de sélectionner ces échantillons de façon qu'ils puissent, chacun, être considérés comme pris au hasard parmi les personnes présentant la pathologie à traiter, l'échantillon traité (respectivement l'échantillon témoin) peut alors être envisagé comme issu d'une population virtuelle de patients traités (respectivement de patients non traités). On se place ici dans une situation de test unilatérale, cherchant à voir si l'on peut décider que le traitement est efficace, **en moyenne,** selon un critère quantitatif approprié. Dans la mesure où le traitement ne peut qu'avoir soit aucun effet soit un effet nécessairement bénéfique, on pourrait se restreindre à une hypothèse nulle ponctuelle, par exemple $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 < 0$, comme on le fait dans certains ouvrages. Cette restriction, comme on l'a vu en particulier en section 9.4.2, ne modifiant pas le niveau du test ni sa puissance par rapport

au test $H_0 : \mu_1 - \mu_2 \leq 0$, nous en resterons à cette dernière hypothèse nulle qui est plus générale.

C'est à ces tests de comparaison des moyennes que l'on doit l'expression «hypothèse nulle». Ceci est également le cas pour l'expression courante du praticien qui parle de «test significatif au niveau α» lorsque l'hypothèse nulle peut être rejetée à ce niveau. En effet ceci se dit par extension de l'idée qu'une différence de deux moyennes empiriques est ou n'est pas **statistiquement significative** selon que le test d'égalité des moyennes théoriques est rejeté ou accepté.

Comme pour la construction d'un intervalle de confiance vue en section 7.4.3, il n'existe pas de méthode exacte dans le cas général où $\sigma_1^2 \neq \sigma_2^2$, mais une procédure asymptotique que nous présenterons par la suite. On suppose donc pour l'heure que **les deux lois ont même variance** σ^2. Dès lors notre point de départ est le même que celui de la section 7.4.3 avec le résultat général suivant :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t(n_1 + n_2 - 2)$$

$$\text{où} \quad S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

est un estimateur sans biais de la variance commune σ^2.

Référons-nous tout d'abord au cas bilatéral : $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$. Sous H_0 la statistique $(\overline{X}_1 - \overline{X}_2)/S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ suit donc une loi $t(n_1 + n_2 - 2)$ ce qui permet de définir une région de rejet, pour un test de risque α, par :

$$\frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \notin [-t_{1-\alpha/2}^{(n_1+n_2-2)}, t_{1-\alpha/2}^{(n_1+n_2-2)}] \text{ où } s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Pour la situation unilatérale $H_0 : \mu_1 \leq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$, on voit intuitivement qu'il faut rejeter uniquement dans $]\, t_{1-\alpha}^{(n_1+n_2-2)}, +\infty[$. Au-delà de l'intuition, on peut montrer que les résultats propres au test sur la moyenne d'un seul échantillon s'étendent à la situation présente. Notamment, avec une telle région de rejet, le risque α maximal est atteint pour $\mu_1 = \mu_2$ et le test proposé est donc bien de niveau α. Pour $H_0 : \mu_1 \geq \mu_2$ vs. $H_1 : \mu_1 < \mu_2$ le rejet se fait dans $]-\infty, -t_{1-\alpha}^{(n_1+n_2-2)}[$.

Ces tests sont UPP-sans biais mais la détermination même de la fonction puissance n'est pas formellement possible. Par ailleurs, l'approche par le rapport de vraisemblance est équivalente (voir exercices).

Pour ce qui concerne la robustesse de ce test vis-à-vis des conditions de lois gaussiennes et d'égalité des variances, les considérations de la section 7.4.3 pour

l'intervalle de confiance restent valables. Rappelons brièvement que, comme pour un seul échantillon, les conditions de distributions gaussiennes ne sont pas cruciales et, d'autre part, que celle d'égalité des variances peut être assouplie dans la mesure où les tailles d'échantillons restent du même ordre. A la lumière des développements de ce chapitre, il est intéressant de revenir sur la pratique mentionnée en section 7.4.3, consistant à effectuer le test de la section suivante pour décider de l'égalité des variances. Cette approche, pour être rassurante pour le praticien, n'offre pas une garantie absolue de l'applicabilité du *test de Student* ci-dessus. En effet nous savons maintenant que l'acceptation d'une hypothèse ne signifie pas qu'elle soit vraie, le risque d'erreur de deuxième espèce n'étant pas contrôlé.

Si **les deux lois mères n'ont pas même variance** on peut utiliser, comme pour l'intervalle de confiance, le fait que, pour de grands échantillons (n_1 et n_2 supérieurs à 100) :

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1)\,.$$

Ainsi, à un niveau approximativement égal à α, on a, par exemple dans le cas bilatéral, la règle de rejet :

$$\frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \notin [-z_{1-\alpha/2}\,,\, z_{1-\alpha/2}]\,,$$

ces quantiles étant lus sur la loi $\mathcal{N}(0\,;1)$. Pour des tailles d'échantillons plus faibles il existe des formules d'approximation dont l'usage n'est pas très répandu.

On peut étendre les résultats ci-dessus au test d'hypothèses nulles du type $H_0 : \mu_1 - \mu_2 = \Delta_0$, $H_0 : \mu_1 - \mu_2 \leq \Delta_0$ ou $H_0 : \mu_1 - \mu_2 \geq \Delta_0$. Il suffit pour cela de retrancher la valeur Δ_0 à celle de $\overline{x}_1 - \overline{x}_2$. En revanche le test bilatéral $H_0 : |\mu_1 - \mu_2| \leq \Delta_0$ vs. $|\mu_1 - \mu_2| > \Delta_0$ pose des difficultés majeures. On peut toutefois recourir à un test conservateur simple d'une façon tout à fait analogue à celle exposée en fin de section 9.7.1 pour une moyenne.

Cas d'échantillons appariés

Cette situation a été décrite en section 7.4.3. Comme pour l'intervalle de confiance on se ramène au cas d'un seul échantillon en étudiant la série des différences entre paires. Soit $\overline{d}$ et s_d respectivement la moyenne et l'écart-type observés pour ces paires, le test se fonde sur la réalisation :

$$\frac{\overline{d}}{s_d/\sqrt{n}}$$

où n est le nombre de paires. Les quantiles définissant les valeurs critiques sont à lire sur une loi de Student à $n-1$ degrés de liberté. Toutes les considérations de la section 9.7.1 restent valables.

9.7.4 Tests de comparaison des variances de deux lois de Gauss

Avec les mêmes notations qu'à la section précédente on a établi, en section 5.5, le résultat général suivant :

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F(n_1-1, n_2-1).$$

Ceci est particulièrement approprié pour tester l'égalité des variances selon $H_0 : \sigma_1^2/\sigma_2^2 = 1$ vs. $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$. En effet, sous H_0 la statistique S_1^2/S_2^2 suit la loi de Fisher mentionnée ci-dessus. La règle de rejet au niveau α sera donc :

$$\frac{s_1^2}{s_2^2} \notin [F_{\alpha/2}^{(n_1-1,n_2-1)}, F_{1-\alpha/2}^{(n_1-1,n_2-1)}].$$

Rappelons pour l'usage des tables que le quantile $F_{\alpha/2}^{(n_1-1,n_2-1)}$ est égal à $1/\ F_{1-\alpha/2}^{(n_2-1,n_1-1)}$.

Ce résultat étant peu robuste par rapport aux conditions gaussiennes, son intérêt est limité.

9.7.5 Tests sur le paramètre p d'une loi de Bernoulli (ou test sur une proportion)

Les applications de ces tests sont multiples dès lors que l'on veut étudier un caractère binaire dans une population. Citons notamment le contrôle de qualité où l'on souhaite vérifier si le taux de produits défectueux ne dépasse pas une valeur donnée.

Le test d'une hypothèse unilatérale a déjà été abordé à la suite de l'exemple 9.6. On a vu que le test UPP repose sur le nombre total de succès observé. Pour $H_0 : p \leq p_0$ vs. $H_1 : p > p_0$, par exemple, on rejette si ce nombre est trop élevé et la valeur critique se lit sur la loi $\mathcal{B}(n, p_0)$. Étant donné le caractère discret de cette loi, le test n'est UPP que dans la mesure où l'on introduit une règle de rejet randomisée afin d'obtenir un risque exactement égal à α (voir note 9.1). Pour l'hypothèse bilatérale $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$ on a un test UPP-sans biais en choisissant deux quantiles extrêmes sur la loi $\mathcal{B}(n, p_0)$ vérifiant une contrainte difficile à mettre en pratique. On préfère donc utiliser des quantiles correspondant au plus proche, et de façon conservatrice, à une équirépartition sur chaque extrémité de cette loi. Cette approche a les mêmes fondements que l'approche des intervalles de confiance par la méthode des quantiles présentée en section 7.5 .

On peut également utiliser une approximation gaussienne comme pour l'intervalle de confiance. La condition d'applicabilité que nous avions retenue est

$n\widehat{p}(1-\widehat{p}) > 12$. Mais ici, sous H_0, la valeur de p est donnée et nous pouvons prendre la condition $np_0 \geq 5$ et $n(1-p_0) \geq 5$ indiquée en section 5.8.3 (toutefois, étant donné que l'on veut une bonne précision sur des quantiles aux extrémités de la loi ces conditions sont souvent renforcées, par exemple en exigeant $np_0(1-p_0) \geq 10$). Alors la statistique $\widehat{P}$, proportion de succès dans l'échantillon, est telle que, sous H_0 :

$$\frac{\widehat{P} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1)\,.$$

D'où l'on déduit, par exemple pour le cas bilatéral, la règle de rejet :

$$\frac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} \notin [-z_{1-\alpha/2}\,,\, z_{1-\alpha/2}]$$

où $\widehat{p}$ est la proportion de succès dans l'échantillon réalisé.

Notons que cette approche est en correspondance avec celle de la procédure d'IC conduisant à la résolution d'une inégalité du second degré (voir section 7.4.5), mais pas avec celle de la formule classique où $p(1-p)$ est estimé par $\widehat{p}(1-\widehat{p})$, ce qui n'est pas nécessaire ici puisque p est spécifié sous H_0. Cependant il est intéressant de noter qu'en substituant $\widehat{p}(1-\widehat{p})$ à $p_0(1-p_0)$ la statistique ci-dessus est quasiment identique à la statistique de Student pour le test sur une moyenne. En effet grâce au codage 1/0, la moyenne de l'échantillon $x_1, x_2, \cdots, x_n$ est $\widehat{p}$ et la variance empirique est (en remarquant que $\sum_{i=1}^n x_i^2$ est le nombre de succès) :

$$\widetilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \widehat{p}^2 = \widehat{p} - \widehat{p}^2 = \widehat{p}(1-\widehat{p}).$$

Il ne manque qu'un facteur $\sqrt{n/(n-1)}$ pour retrouver l'écart-type de l'échantillon plutôt que l'écart-type empirique.

On a donc avantage, dans un fichier de données, à utiliser ce codage 1/0 pour une variable binaire car certains logiciels rudimentaires ne proposent que le test de Student (le fait que la P-valeur soit calculée sur la loi de Student plutôt que sur la loi de Gauss ne posant pas de problème puisque cela va dans le sens conservateur).

On peut voir aussi le test par approximation gaussienne comme une application du résultat asymptotique de l'estimateur du maximum de vraisemblance $\widehat{P}$ pour p, mentionnée en section 9.6, car l'information de Fisher $I(p_0)$ est égale à $[\,p_0(1-p_0)]^{-1}$ pour la loi de Bernoulli.

Par ailleurs, pour le cas bilatéral, le test du RVG ne donne ni le test exact ni le test approché ci-dessus mais on verra qu'il est asymptotiquement équivalent au test du khi-deux, lui-même identique au test approché (voir sections 10.1.3 et 10.1.4).

9.7.6 Tests de comparaison des paramètres de deux lois de Bernoulli (comparaison de deux proportions)

La comparaison de deux proportions à partir de deux échantillons indépendants est très fréquente, au même titre que la comparaison de deux moyennes. Elle s'applique également à l'étude de l'efficacité d'un traitement par rapport à un autre (ou en l'absence de traitement avec un échantillon témoin) lorsque cette efficacité est évaluée par un critère binaire, par exemple la guérison ou non guérison d'un patient. Pour les mêmes raisons que celles invoquées en section 9.7.3, dans la mesure où le traitement ne saurait avoir d'effet négatif par rapport à une absence de traitement, on pourrait se restreindre à une hypothèse nulle ponctuelle $H_0 : p_1 - p_2 = 0$ avec une hypothèse alternative unilatérale. Ici également on conclura en déclarant que la différence est ou n'est pas significative à un niveau donné.

Soit S_1 et S_2 les statistiques exhaustives minimales des nombres de succès parmi les n_1 et n_2 observations respectives de chaque échantillon. On obtient un test UPP-sans biais en utilisant la loi conditionnelle de S_1 (ou de S_2) sachant $S_1 + S_2$. Montrons tout d'abord que, sous $H_0 : p_1 = p_2$, cette loi est une loi hypergéométrique et notons $p = p_1 = p_2$ sous cette hypothèse nulle.

Rappelons (voir section 4.1.3) que si $S_1 \rightsquigarrow \mathcal{B}(n_1, p)$ et $S_2 \rightsquigarrow \mathcal{B}(n_2, p)$, alors $S_1 + S_2 \rightsquigarrow \mathcal{B}(n_1 + n_2, p)$. On a donc :

$$
\begin{aligned}
P(S_1 = x \,|\, S_1 + S_2 = t) &= \frac{P(S_1 = x,\, S_2 = t - x)}{P(S_1 + S_2 = t)} \\
&= \frac{\binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{t-x} p^{t-x} (1-p)^{n_2 - t + x}}{\binom{n_1 + n_2}{t} p^t (1-p)^{n_1 + n_2 - t}} \\
&= \frac{\binom{n_1}{x}\binom{n_2}{t-x}}{\binom{n_1 + n_2}{t}}
\end{aligned}
$$

qui est le terme général de la fonction de probabilité de la loi $\mathcal{H}(n_1 + n_2, t, n_1)$ selon les notations de la section 4.1.5.

Le test UPP-sans biais est défini, par exemple dans le cas bilatéral, par une région de rejet de la forme $s_1 \notin [c_{\alpha_1}, c_{1-\alpha_2}]$ où c_{α_1} et $c_{1-\alpha_2}$ sont des quantiles d'ordres α_1 et $1-\alpha_2$ tels que $\alpha_1 + \alpha_2 = \alpha$, issus de la loi $\mathcal{H}(n_1+n_2, s_1+s_2, n_1)$ où s_1 et s_2 sont les nombres de succès observés sur les échantillons réalisés. On montrera en section 10.3.2 (note 10.3) qu'un test défini, comme celui-ci, conditionnellement à une statistique exhaustive sous H_0 (ce qu'est $S_1 + S_2$ ici) est légitime au sens où il s'agit bien d'un test de niveau α non conditionnellement.

Les calculs à la main sont très fastidieux mais certains logiciels donnent la P-valeur pour ce test. On peut également utiliser le *test exact de Fisher* pour l'indépendance de deux v.a. binaires dont les calculs sont parfaitement identiques (voir section 10.3.2 et exemple 10.1 pour la mise en oeuvre).

La puissance est difficilement accessible pour une alternative quelconque (toutefois elle peut être calculée en fonction du *rapport des chances* - en anglais : *odds ratio* - $[\, p_2/(1-p_2)]/[\, p_1/(1-p_1)]$ dont dépend la loi conditionnelle de S_1 sachant $S_1 + S_2$ sous H_1).

Dès lors que l'approximation gaussienne vaut pour chaque échantillon on utilise une formulation approchée simple. Les conditions de validité que nous retenons sont $n_1\widehat{p}_1(1-\widehat{p}_1) > 12$ et $n_2\widehat{p}_2(1-\widehat{p}_2) > 12$ où $\widehat{p}_1$ et $\widehat{p}_2$ sont les proportions de succès observées. On a établi pour l'IC correspondant (voir section 7.4.6) la loi approximative de $\widehat{P}_1 - \widehat{P}_2$:

$$\widehat{P}_1 - \widehat{P}_2 \underset{approx}{\rightsquigarrow} \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right),$$

ce qui donne sous H_0 :

$$\widehat{P}_1 - \widehat{P}_2 \underset{approx}{\rightsquigarrow} \mathcal{N}\left(0\,;\, p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})\right).$$

Toutefois la valeur de p est inconnue et doit être estimée par la proportion de succès dans les deux échantillons fusionnés, i.e.

$$\widehat{p} = \frac{s_1 + s_2}{n_1 + n_2} = \frac{n_1\widehat{p}_1 + n_2\widehat{p}_2}{n_1 + n_2}$$

d'où, finalement, la région de rejet (cas bilatéral) :

$$\frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \notin [-z_{1-\alpha/2}\,,\, z_{1-\alpha/2}].$$

On montrera dans le cadre plus général de la comparaison de lois multinomiales (voir section 10.2) que ce test bilatéral est identique au test classique du khi-deux.

Par ailleurs, pour le cas bilatéral, le test du RVG ne donne ni le test exact ni ce test approché mais on verra (section 10.2) que la forme de la statistique du rapport de vraisemblance est asymptotiquement équivalente à celle de la statistique du khi-deux, donc de la statistique ci-dessus obtenue par approximation gaussienne.

Notons bien que ces tests ne sont valables que pour des échantillons indépendants. Pour des **échantillons appariés** citons simplement le *test de McNemar* (la construction de ce test est proposée dans un exercice du chapitre 10). Il s'appliquera dans les enquêtes, par exemple, pour tester s'il y a une évolution significative dans la réponse à une certaine modalité d'une question pour un même échantillon réinterrogé après un certain laps de temps (situation de mesures répétées).

Note 9.6 Comparaisons de proportions au sein d'un même échantillon

On peut vouloir tester l'égalité de deux proportions entre deux sous-échantillons. Par exemple tester que le pourcentage de réponses à une modalité d'une question dans une enquête est le même pour les femmes et pour les hommes. Nous ne sommes plus dans le schéma précédent du fait que n_1 et n_2 (fréquences des femmes et fréquences des hommes) ne sont plus fixées a priori mais sont des variables aléatoires. En réalité, il s'agit ici d'un test d'indépendance entre la variable sexe et le choix ou non de la modalité de réponse. Ce test sera présenté en section 10.3 où l'on verra que statistique de test et région critique seront, en fait, identiques à celles des tests présentés ci-dessus, que ce soit pour le test exact ou pour le test par approximation gaussienne.

On peut encore vouloir comparer les pourcentages de réponses p_1 et p_2 pour deux modalités distinctes d'une même question. Par exemple, dans un sondage aléatoire sur les intentions de vote, voir si les pourcentages obtenus par deux candidats diffèrent ou non de façon statistiquement significative. Ici on est dans le cas d'une loi multinomiale (il y a plusieurs modalités de réponse à la question) et les fréquences observées ne sont pas indépendantes. Soit N_1 et N_2 ces fréquences aléatoires, nous avons vu en section 4.1.6 que $cov(N_1, N_2) = -np_1p_2$, d'où pour les proportions observées $\widehat{P}_1 = N_1/n$ et $\widehat{P}_2 = N_2/n$: $cov(\widehat{P}_1, \widehat{P}_2) = -p_1p_2/n$. La variance de $\widehat{P}_1 - \widehat{P}_2$ est donc égale à :

$$V(\widehat{P}_1) + V(\widehat{P}_2) - 2cov(\widehat{P}_1, \widehat{P}_2) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n} + \frac{2p_1p_2}{n}.$$

Sous $H_0 : p_1 = p_2 = p$ on a, en admettant que la loi de $\widehat{P}_1 - \widehat{P}_2$ soit toujours approximativement gaussienne :

$$\widehat{P}_1 - \widehat{P}_2 \underset{approx}{\rightsquigarrow} \mathcal{N}\left(0; \frac{2p}{n}\right)$$

où p doit être estimé, un estimateur naturel étant $\widehat{P} = (N_1 + N_2)/2n$ qui est sans biais. Dans le cas bilatéral on rejettera H_0 si (cas bilatéral) :

$$\frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{2\widehat{p}}{n}}} \notin [-z_{1-\alpha/2}\,,\, z_{1-\alpha/2}].$$

9.7.7 Test sur la corrélation dans un couple gaussien

Soit le couple aléatoire (X, Y) de loi gaussienne bivariée. Pour le cas général d'un vecteur gaussien dans $\mathbb{R}^p$ on a vu en section 3.9 que la densité conjointe de ses composantes au point $(x_1, x_2, \cdots, x_p) \in \mathbb{R}^p$ est :

$$f_{\mathbf{X}}(x_1, x_2, \cdots, x_p) = \frac{1}{(2\pi)^{p/2}(\det \boldsymbol{\Sigma})^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \ \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}$$

où $\boldsymbol{\mu}$ est le vecteur des moyennes et $\boldsymbol{\Sigma}$ est la matrice des variances-covariances. Pour un couple ($p = 2$) on a :

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \text{et} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

où $\rho\sigma_1\sigma_2$ est la covariance des deux composantes et ρ est leur corrélation linéaire (voir définition 3.6). La loi de (X, Y) dépend donc de cinq paramètres et la densité conjointe s'écrit :

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

On considère un échantillon de taille n : $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ issu de cette loi dont les cinq paramètres sont inconnus et l'on souhaite tester l'hypothèse nulle d'indépendance des deux composantes. Comme il a été vu en section 3.6 ceci équivaut à tester que la corrélation est nulle, soit :

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0\,.$$

A partir de l'expression $\prod_{i=1}^n f_{X,Y}(x_i, y_i)$ de la densité conjointe de l'échantillon on peut établir l'expression du rapport de vraisemblance généralisé et montrer (non sans difficultés) qu'elle ne dépend des observations qu'à travers la réalisation r de la corrélation linéaire empirique définie en section 5.2 :

$$R = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}\ \sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}\,,$$

la région de rejet $\lambda < k$ étant équivalente à $|r| > k'$, ce qui semble naturel. On utilise plutôt comme statistique de test la fonction croissante de R :

$$T = \frac{\sqrt{n-2}\,R}{\sqrt{1-R^2}}$$

qui offre l'avantage de suivre simplement une loi de Student à $n-2$ degrés de liberté sous H_0 (une démonstration sera donnée en section 11.2.6). On rejette donc H_0 au niveau α si la réalisation t de T tombe en dehors de l'intervalle $[-t^{(n-2)}_{1-\alpha/2}\,,\, t^{(n-2)}_{1-\alpha/2}]$.

Note 9.7 A partir de la loi de T on peut, par simple changement de variable, établir la loi de R sous H_0 et montrer que sa fonction de densité est :

$$f_R(r) = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-2}{2})}(1-r^2)^{(n-4)/2} \quad \text{si } r \in [-1,1]\,.$$

Cette forme n'est pas sans rappeler celle d'une loi bêta. En fait on peut voir aisément que le *coefficient de détermination* R^2 suit une loi *Beta(0,(n-4)/2)* et il est équivalent de fonder le test sur la réalisation r^2 de ce coefficient avec un rejet pour $r^2 > c_{1-\alpha}$ où $c_{1-\alpha}$ est un quantile de cette loi.

On peut également déterminer l'expression de la densité de R pour ρ quelconque et établir la fonction puissance du test. Signalons que la loi de R ne dépend que du paramètre ρ, que $E(R) = \rho + O(\frac{1}{n})$ et que $V(R) = \frac{(1-\rho^2)^2}{n} + o(\frac{1}{n})$. R est donc un estimateur biaisé de r pour n fini, mais il est asymptotiquement sans biais et convergent en moyenne quadratique (et presque sûrement en tant que fonction continue de moments empiriques). On démontrera également en section 11.2.6 que R est l'estimateur du maximum de vraisemblance de ρ.

Fisher a établi un résultat asymptotique pour ρ quelconque :

$$\frac{1}{2}\ln\frac{1+R}{1-R} \underset{approx}{\rightsquigarrow} \mathcal{N}(\frac{1}{2}\ln\frac{1+\rho}{1-\rho}\,,\,\frac{1}{n-3})$$

qui autorise un test approximatif pour une hypothèse générale $H_0 : \rho = \rho_0$ (bilatérale ou unilatérale).

Pour ces tests l'hypothèse d'une loi gaussienne pour le couple est essentielle et les résultats obtenus sont assez peu robustes. Si cette condition est douteuse on pourra se tourner vers une procédure non paramétrique telle que le test sur la corrélation des rangs de Spearman (voir section 10.5.5). Rappelons qu'en dehors de la loi de Gauss l'hypothèse nulle ne signifie qu'une absence de corrélation et non pas l'indépendance des deux composantes du couple. Pour tester l'indépendance on pourra recourir au test concernant les variables catégorielles (voir section 10.3) en découpant X et Y en classes.

9.8 Dualité entre tests et intervalles de confiance

La présentation des tests usuels a permis de voir que ces procédures et celles utilisées pour la construction d'intervalles de confiance sont très voisines. Nous allons voir que l'on peut même établir une sorte d'équivalence entre celles-ci. Montrons-le sur un exemple avant de passer au cas général.

Considérons le test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ pour une loi mère $\mathcal{N}(\mu, \sigma^2)$ où (μ, σ^2) est inconnu. On accepte H_0 au niveau α si et seulement si :

$$-t_{1-\alpha/2}^{(n-1)} < \frac{\overline{x} - \mu_0}{s/\sqrt{n}} < t_{1-\alpha/2}^{(n-1)},$$

ce que l'on peut écrire de façon équivalente :

$$\overline{x} - t_{1-\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}} < \mu_0 < \overline{x} + t_{1-\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}.$$

Ainsi pour qu'une valeur de μ hypothétique soit acceptée il faut et il suffit qu'elle soit dans l'intervalle $[\overline{x} - t_{1-\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}, \overline{x} + t_{1-\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}]$, c'est-à-dire qu'elle soit comprise dans l'intervalle de confiance de niveau $1 - \alpha$ pour la moyenne inconnue μ. Il y a donc équivalence pour μ entre le fait de prendre une valeur acceptée dans le test de niveau α et le fait d'être situé dans l'intervalle de confiance de niveau $1 - \alpha$. On peut donc voir aussi l'IC comme l'ensemble des valeurs acceptées par le test. Essayons de formaliser cela dans la généralité.

IC dérivé d'un test

Soit pour l'hypothèse nulle $H_0 : \theta = \theta_0$ un test de niveau α défini par la région d'acceptation $\mathbb{A}(\theta_0) \subset \mathbb{R}^n$ donc telle que $P_{\theta_0}(\mathbb{A}(\theta_0)) = 1 - \alpha$. Faisant varier θ_0 dans Θ, on peut ainsi construire, pour chaque valeur de $\theta \in \Theta$, une région d'acceptation qui dépend de cette valeur, notée $\mathbb{A}(\theta)$ et telle que $P_\theta(\mathbb{A}(\theta)) = 1 - \alpha$.

Soit maintenant **une région de** Θ construite de la façon suivante sur la base d'une réalisation $(x_1, x_2, \cdots, x_n)$. On considère chaque valeur de θ dans Θ et l'on inclut cette valeur dans la région si et seulement si $(x_1, x_2, \cdots, x_n) \in \mathbb{A}(\theta)$. Passant à l'univers des réalisations possibles symbolisé par $(X_1, X_2, \cdots, X_n)$, la région ainsi définie devient aléatoire. Or, pour un θ donné, la probabilité que $(X_1, X_2, \cdots, X_n)$ appartienne à $\mathbb{A}(\theta)$ est égale à $1 - \alpha$ par construction même de $\mathbb{A}(\theta)$. Comme il y a identité entre cet événement et le fait d'inclure cette valeur de θ dans la région de Θ, la probabilité que θ soit compris dans cette région (aléatoire) est égale à $1 - \alpha$. Ainsi on a construit une procédure de région de confiance de niveau $1 - \alpha$ pour le paramètre inconnu θ. Quand $\Theta \subseteq \mathbb{R}$ cette région sera généralement un intervalle quel que soit θ et on aura une procédure d'intervalle de confiance.

On peut montrer que les propriétés d'optimalité du test se transfèrent à la procédure d'IC. Ainsi un test UPP donnera une procédure uniformément

plus précise (voir section 7.7). Un test sans biais impliquera que la procédure fournira une probabilité plus faible d'inclure une fausse valeur de θ que d'inclure la vraie valeur. En général si l'alternative est bilatérale la région d'acceptation est un intervalle à bornes finies et il en est de même pour l'IC. Une alternative unilatérale conduira à un intervalle de confiance ouvert sur l'infini pour un côté.

Dans le cas discret la dualité n'est valable qu'en «randomisant» les bornes de l'intervalle. Il sera toutefois plus simple de partir de régions d'acceptation conservatrices pour aboutir à un IC de niveau de confiance supérieur ou égal à $1-\alpha$.

Pour illustrer l'intérêt de cette dualité prenons la méthode du test par le rapport de vraisemblance (généralisé) avec $\Theta \subseteq \mathbb{R}$. En vertu de ses propriétés asymptotiques on a une région d'acceptation au niveau approximatif α définie par (voir note 9.4) :

$$\begin{aligned} -2\ln\lambda(x_1,x_2,...,x_n) &\leq \chi^{2\ (1)}_{1-\alpha} \\ \text{ou} \qquad \lambda(x_1,x_2,...,x_n) &\geq e^{-\frac{1}{2}\chi^{2\ (1)}_{1-\alpha}} \\ \text{ou} \quad f(\theta;x_1,x_2,...,x_n) &\geq e^{-\frac{1}{2}\chi^{2\ (1)}_{1-\alpha}}\ f(\widehat{\theta}_n;x_1,x_2,...,x_n). \end{aligned}$$

Pour un échantillon réalisé $x_1, x_2, \cdots, x_n$ on en déduit une région de confiance de niveau $1-\alpha$ en considérant l'ensemble des valeurs de θ vérifiant cette dernière inégalité, cette région étant généralement un intervalle étant donné les propriétés du RVG. Ainsi avec α =0,05 , on obtient un IC de niveau approximatif 0,95 contenant les valeurs de θ pour lesquelles la densité (ou la fonction de probabilité) conjointe aux valeurs observées $x_1, x_2, \cdots, x_n$ n'est pas inférieure à 0,147 fois leur densité maximale $f(\widehat{\theta}_n; x_1, x_2, ..., x_n)$ atteinte au maximum de vraisemblance $\widehat{\theta}_n$ (voir note 9.4). Ceci est représenté sur la figure 9.2.

D'un point de vue numérique on voit qu'il suffit de connaître l'expression de la densité (ou de la fonction de probabilité) conjointe des observations pour donner un intervalle de confiance approximatif sur le paramètre inconnu.

Test dérivé d'un IC

On peut également envisager une démarche inverse permettant de déboucher sur un test à partir d'une procédure d'IC. Soit une famille d'intervalles $[t_1(x_1,x_2,...,x_n)\,, t_2(x_1,x_2,...,x_n)]$ définie pour toute réalisation $(x_1,x_2,...,x_n)$, où $t_1(x_1,x_2,...,x_n)$ et $t_2(x_1,x_2,...,x_n)$ sont à valeurs dans $\Theta \subseteq \mathbb{R}$, issue d'une procédure d'IC de niveau $1-\alpha$.

Pour tout θ_0 on peut définir un test $H_0 : \theta = \theta_0$ consistant à accepter H_0 si et seulement si θ_0 appartient à $[t_1(x_1,x_2,...,x_n), t_2(x_1,x_2,...,x_n)]$. Par construction de la procédure d'IC, pour toute valeur de θ on a :

$$P_\theta(\theta \in [t_1(X_1,X_2,\cdots,X_n), t_2(X_1,X_2,\cdots,X_n)]\,) = 1-\alpha.$$

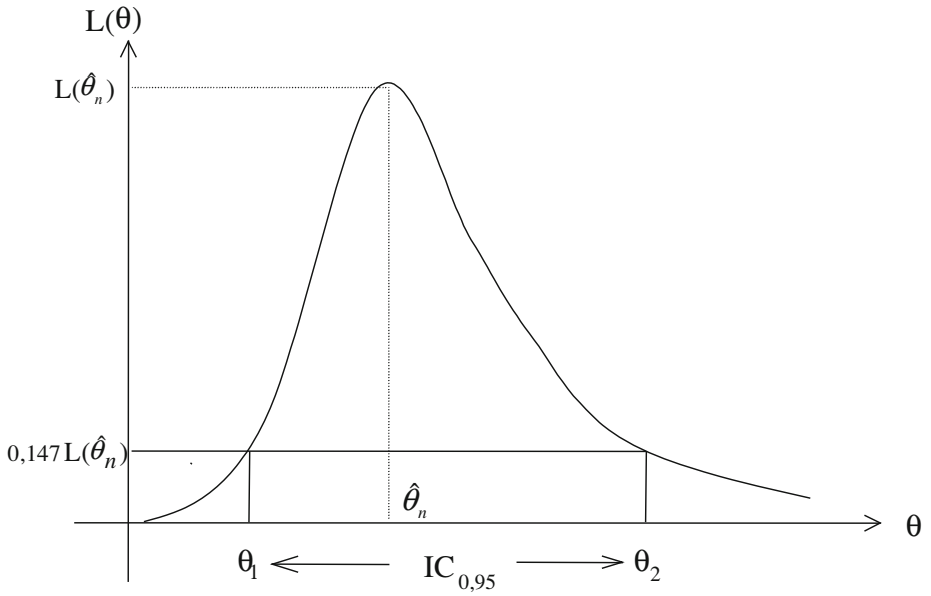


Figure 9.2 - Intervalle de confiance dérivé du test du rapport de vraisemblance généralisé.

Ceci est vrai en particulier pour θ_0 et la probabilité d'accepter cette valeur sous H_0 est aussi égale à $1-\alpha$. On a bien un test de niveau α.

Ainsi, par exemple, on peut recourir à la méthode des quantiles (voir section 7.5), facile à mettre en oeuvre, pour tester une valeur de θ : il suffit de voir si cette valeur est ou non à l'intérieur des limites de confiance. En particulier les abaques de confiance pour le paramètre p de la loi de Bernoulli ou λ de la loi de Poisson peuvent être utilisés dans cette optique.

Nous concluons en disant qu'un intervalle de confiance donne une information plus riche qu'un simple test car il indique l'ensemble des valeurs qui seraient acceptables via le test dual.

Pour approfondir la théorie des tests on pourra consulter l'ouvrage de référence de Lehmann (1986) ou celui de Shao (1999). Par ailleurs on trouvera dans le livre de Saporta (1990) un vaste éventail de méthodes où s'appliquent les tests les plus divers.

9.9 Exercices

Exercice 9.1 Soit $X_1, X_2, \cdots, X_n$ issus d'une loi $\mathcal{E}(\lambda)$. On souhaite tester $H_0 : \lambda = 1/2$ vs. $H_1 : \lambda = 1$. Quelle est la région de rejet au niveau 0,05 pour

le test du RV simple ?

Aide : on pourra utiliser le fait qu'une loi $\Gamma(n, 1/2)$ est une loi $\chi^2(2n)$.

Exercice 9.2 Soit un échantillon aléatoire X_1, X_2, ... , X_n issu d'une loi géométrique de paramètre inconnu p dont nous rappelons la fonction de probabilité :

$$f(x;p) = p(1-p)^x \; ; \; x = 0, 1, 2, ...$$

Expliquer pourquoi $\sum_{i=1}^n X_i$ suit une loi binomiale négative de paramètres n et p. Soit à tester $H_0 : p = \dfrac{1}{3}$ versus $H_1 : p = \dfrac{2}{3}$.
Montrer que la région critique pour le test fondé sur le rapport de vraisemblance est de la forme $\sum_{i=1}^n x_i < k$.
Donner la règle de décision pour $n = 4$ et $\alpha = 0{,}05$.
Quelle est la puissance de ce test ?

Exercice 9.3 Soit $X_1, X_2, \cdots, X_n$ issus de la loi $\mathcal{N}(\mu\,, 1)$. Pour tester H_0 : $\mu \leq 5$ vs. $H_1 : \mu > 5$ on adopte la région de rejet :

$$\{\,(x_1, x_2, \cdots, x_n) \mid \; \overline{x} > 5 + \frac{1}{\sqrt{n}}\}.$$

Quel est le risque de première espèce de ce test ? Déterminer sa fonction puissance.

Exercice 9.4 Soit la famille de lois de Pareto de paramètres a connu et θ inconnu (voir section 4.2.6). Montrer qu'elle est à RV monotone. En déduire le test UPP pour $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$.
Application : pour $a = 1$ construire le test de l'hypothèse nulle : la moyenne de la loi est inférieure ou égale à 2. (aide : la moyenne est $\theta a/(\theta - 1)$).
Exprimer la valeur critique pour $\alpha = 0{,}05$.

Aide : montrer que $\ln X$ suit une loi exponentielle.

Exercice 9.5 Soit la famille de lois $\mathcal{U}[0, \theta]$ de paramètre θ inconnu. Montrer qu'elle est à rapport de vraisemblance non croissant. En déduire un test UPP de niveau $\alpha > 0$ pour des alternatives unilatérales sur θ.

Exercice 9.6 Soit la situation du tirage sans remise de n individus dans une population de N individus dont M ont un certain caractère, M étant inconnu. On considère le nombre $X \rightsquigarrow \mathcal{H}(N, M, n)$ d'individus ayant ce caractère dans un échantillon de taille n.
Soit l'hypothèse $H_0 : M \geq M_0$ vs. $H_1 : M < M_0$, montrer que le test avec région de rejet $x \leq c_\alpha$, où c_α est le quantile d'ordre α sur la loi $\mathcal{H}(N, M_0, n)$, est UPP.

Aide : montrer que la famille des lois hypergéométriques avec M inconnu est à rapport de vraisemblance croissant. Pour cela il suffira de montrer que $L(M+1; x)/L(M; x)$ est une fonction croissante de x.

Exercice 9.7 Soit la famille de lois exponentielles $\mathcal{E}(\lambda)$. Sur la base d'une seule observation donner, en accord avec la note 9.3, les deux équations conduisant au choix des valeurs critiques pour le test UPP parmi les tests sans biais de $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. Montrer que la région de rejet n'est pas répartie de façon égale selon un risque de première espèce $\alpha/2$ sur chaque extrémité (on pourra constater que le déséquilibre peut être très prononcé sur un cas particulier en se donnant λ_0 et α, et en recourant à un logiciel de résolution d'équations).

Exercice 9.8 Soit une loi mère $\mathcal{N}(\mu, \sigma^2)$, où μ est inconnu mais σ^2 est connu, et la situation de test $H_0 : \mu_1 \leq \mu \leq \mu_2$ vs. $H_1 : \mu < \mu_1$ ou $\mu > \mu_2$. Déterminer la forme de la région de rejet pour le test du RVG au niveau α. En admettant que le risque de première espèce est maximal en $\mu = \mu_1$ et $\mu = \mu_2$, et en prenant naturellement des valeurs critiques symétriques par rapport à $\frac{\mu_1+\mu_2}{2}$ donner l'équation définissant ces valeurs critiques (ceci correspond au test UPP parmi les tests sans biais). Application : résoudre approximativement l'équation pour $\mu_1 = 4$, $\mu_2 = 5$, $\sigma^2 = 1$ et $\alpha = 0,05$. Tracer en un seul graphe la fonction puissance (avec un choix de n) et la variation du risque de première espèce.

Déduire du test précédent le test UPP-sans biais pour $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ et σ^2 connu. Montrer également que, dans ce cas, la loi asymptotique de $-2 \ln \Lambda_n$, où Λ_n est le RVG, est en fait la loi exacte.

Exercice 9.9 Soit à tester $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$ pour le paramètre λ de la loi $\mathcal{E}(\lambda)$ à partir d'un échantillon de taille n. Établir formellement le test du RVG. Application : établir le test pour $\lambda_0 = 1/4$ et $n = 30$ en utilisant la loi asymptotique du RVG.

Exercice 9.10 Soit une loi (de Raleigh) de densité $f(x; a) = 2ax \exp\{-ax^2\}$, $x \geq 0$, $a > 0$. Donner la statistique du RVG pour tester $H_0 : a = 1$ vs. $H_1 : a \neq 1$.

Exercice 9.11 Soit la loi de Pareto de paramètre $a = 2$ et θ inconnu. Donner le test du RVG pour $H_0 : \theta = 3$ vs. $H_1 : \theta \neq 3$ en utilisant la loi asymptotique du RVG. Application : $n = 30$, $\sum_{i=1}^{30} \ln x_i = 31$.

Exercice 9.12 Soit une loi mère $\mathcal{P}(\lambda)$ où λ est inconnu. On veut tester $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. Montrer que la région de rejet pour le test du RVG est de la forme $\overline{x} \notin [c_1, c_2]$ avec une certaine contrainte liant c_1 et c_2.
Application : pour $\lambda_0 = 5$ et $n = 10$ résoudre au plus proche de la solution du RVG en utilisant une région de rejet conservatrice par rapport au niveau 0,05. Calculer le niveau exact de cette règle.
* Construire une règle de rejet «randomisée» selon le principe de la note 9.1.

Exercice 9.13 Montrer que le test du RVG conduit à la même statistique de test que le test classique de Student vu en section 9.7.1 pour le test sur la moyenne d'une loi de Gauss.

Exercices appliqués[3]

Exercice 9.14 Un producteur de pneus envisage de changer la méthode de fabrication. La distribution de la durée de vie de ses pneus traditionnels est connue : moyenne 64 000 km, écart-type 8 000 km ; elle est pratiquement gaussienne. Dix pneus sont fabriqués avec la nouvelle méthode et une moyenne de 67 300 km est constatée. En supposant que la nouvelle fabrication donnerait une distribution à peu près gaussienne et de même variance, testez l'efficacité de la nouvelle méthode au niveau $\alpha = 0,05$. Tracez la fonction puissance de ce test.
(aide : test de $H_0 : \mu \leq \mu_0$)

Exercice 9.15 Une étude approfondie a évalué à 69 800 euros/an le revenu moyen imposable par ménage résidant à Neuilly-sur-Seine. Une enquête est effectuée auprès de 500 ménages pris au hasard, afin de contrôler le résultat de l'étude. Dans l'enquête on trouve une moyenne de 68 750 euros/an avec un écart-type de 10 350 euros/an. Quelle est la P-valeur associée aux résultats du contrôle ?
(aide : test de $H_0 : \mu = \mu_0$)

Exercice 9.16 En un point de captage d'une source on a répété six mesures du taux d'oxygène dissous dans l'eau (en parties par million). On a trouvé :

$$4,92 \quad 5,10 \quad 4,93 \quad 5,02 \quad 5,06 \quad 4,71.$$

La norme en dessous de laquelle on ne doit pas descendre pour la potabilité de l'eau est 5 ppm. Au vu des observations effectuées peut-on avec un faible risque d'erreur affirmer que l'eau n'est pas potable (admettre une distribution quasi-gaussienne des aléas des mesures) ?
(aide : test de $H_0 : \mu \geq \mu_0$)

Exercice 9.17 Un service chargé de traiter des formulaires standard utilise un réseau de dix micro-ordinateurs et une imprimante. Le temps moyen d'attente en impression d'un formulaire est de 42,5 secondes (le temps entre l'envoi de la commande d'impression et la réalisation de l'impression du formulaire).

Dix nouveaux micros et une imprimante sont ajoutés au réseau. Sur trente demandes d'impression dans cette nouvelle configuration on a constaté un temps moyen de 39,0 secondes et un écart-type de 8,2 secondes.

Tester l'hypothèse que le temps moyen d'impression n'a pas été affecté par l'accroissement du réseau.
(aide : test de $H_0 : \mu = \mu_0$)

[3]Un ou deux de ces exercices appliqués sont des adaptations d'emprunts dont nous ne sommes plus en mesure de retrouver la source. Nous nous en excusons auprès des involontaires contributeurs.

Exercice 9.18 On veut tester la précision d'une méthode de mesure d'alcoolémie sur un échantillon sanguin. La précision est définie comme étant égale à deux fois l'écart-type de l'aléa (supposé pratiquement gaussien) de la méthode.

On partage l'échantillon de référence en 6 éprouvettes que l'on soumet à l'analyse d'un laboratoire. Les valeurs trouvées en g/litre sont :

$$1,35 \quad 1,26 \quad 1,48 \quad 1,32 \quad 1,50 \quad 1,44\,.$$

Tester l'hypothèse nulle que la précision est inférieure ou égale à 0,1 g/litre au niveau 0,05. Donner la P-valeur du résultat.
(aide : test de $H_0 : \sigma^2 \leq \sigma_0^2$)

Exercice 9.19 On sait que dans la population générale du nord de l'Italie le pourcentage de prématurés (naissance avant le 8ème mois) est de 4 %. Dans une région du nord de l'Italie contaminée par une pollution chimique on a observé sur les dernières années 72 naissances prématurées sur 1 243 accouchements.

Y-a-t-il lieu, selon la P-valeur constatée, de penser que la proportion de prématurés est plus élevée dans cette région que dans l'ensemble de la population du Nord du pays ? Donnez la fonction puissance du test de niveau 0,01.
(aide : test de $H_0 : p \leq p_0$)

Exercice 9.20 Le fournisseur d'un lot de 100 000 puces affirme que le taux de puces défectueuses ne dépasse pas 4 %. Pour tester cette hypothèse 800 puces prises au hasard sont contrôlées et l'on en trouve 40 défectueuses. Effectuer un test de niveau 0,05.
(aide : test de $H_0 : p \leq p_0$)

Exercice 9.21 Dans une étude on a mesuré le taux de plomb dans le sang (en mg/litre) de 67 enfants tirés au hasard dans les classes primaires d'une ville, dont 32 filles et 35 garçons. Pour les filles on a trouvé une moyenne de 12,50 avec une variance de 3,39. Pour les garçons on a trouvé, respectivement, 12,40 et 3,94. Le taux moyen est-il significativement différent entre garçons et filles ? Quelle hypothèse supplémentaire doit-on faire pour pouvoir répondre à cette question ?
(aide : test de $H_0 : \mu_1 = \mu_2$)

Exercice 9.22 L'an dernier, on a observé sur un échantillon de 29 appartements de 3 pièces situés en ville des dépenses de chauffage égales en moyenne à 325 euros, avec un écart-type égal à 26 euros.

Cette année, pour un nouvel échantillon de 31 appartements de 3 pièces en ville on a trouvé des valeurs respectives de 338 euros et 28 euros. L'hypothèse à laquelle on s'intéresse est qu'il n'y a pas eu d'augmentation des dépenses, en moyenne entre les deux années.

a) En supposant que toutes les conditions nécessaires à la validité du test utilisé sont remplies, effectuer un test de niveau 0,05 pour l'hypothèse ci-dessus.
b) Donner les conditions nécessaires pour que la procédure de test utilisée soit applicable.
(aide : test de $H_0 : \mu_1 \geq \mu_2$)

Exercice 9.23 Pour tester l'efficacité d'un traitement destiné à augmenter le rythme cardiaque, on a mesuré sur 5 individus ce rythme avant et après administration du traitement. Est-il efficace ?

Avant	80	90	70.3	85	63
Après	84	95.5	73.5	86	62

On supposera que le rythme cardiaque se répartit de façon quasi gaussienne pour la population considérée (avant comme après traitement).
(aide : test «apparié» $H_0 : \mu_1 \geq \mu_2$)

Exercice 9.24 Une entreprise qui commercialise des abonnements pour un opérateur de téléphonie mobile, applique un nouveau régime horaire à ses employés. Pour 16 vendeurs pris au hasard, elle comptabilise le nombre d'abonnements vendus le mois précédant l'application du régime et le mois suivant :

vendeur	1	2	3	4	5	6	7	8
Mois précédent	39	28	67	45	28	73	67	53
Mois suivant	43	51	64	35	18	53	66	61
vendeur	9	10	11	12	13	14	15	16
Mois précédent	69	41	52	60	50	46	53	47
Mois suivant	69	43	53	47	34	39	49	56

En se fondant sur cette information, la direction annonce que le nouveau régime provoque une baisse importante des ventes. Cette affirmation est–elle justifiée ? Donner une valeur approchée de la P–valeur du test effectué. On admettra que la loi du nombre de ventes mensuelles est suffisamment proche d'une loi normale.
(aide : test «apparié» $H_0 : \mu_1 \leq \mu_2$)

Exercice 9.25 Un nouveau vaccin contre le paludisme est expérimenté auprès de la population d'une ville d'Afrique.

On prend deux échantillons A et B de 200 personnes chacun. On injecte le vaccin aux individus de l'échantillon A et un placebo à ceux de l'échantillon B. Au bout d'un an on constate que 40 personnes de l'échantillon A ont des accès de palustres et 80 de l'échantillon B. Que dire de l'efficacité du vaccin ?
(aide : test de $H_0 : p_1 \geq p_2$)

Exercice 9.26 Suite à des sondages, l'institut A donne 510 personnes favorables à telle mesure sur 980 personnes interrogées, l'institut B donne 505 favorables sur 1 030.

La différence des estimations de la proportion de personnes favorables est-elle significative ?
(aide : test de $H_0 : p_1 = p_2$)

Chapitre 10

Tests pour variables catégorielles et tests non paramétriques

Dans ce chapitre nous considérons tout d'abord la généralisation des tests des sections 9.7.5 et 9.7.6 concernant des variables de Bernoulli, à des *variables catégorielles*. Les tests sur des variables catégorielles sont voisins dans leur esprit des tests non paramétriques et certains d'entre eux sont effectivement de nature non paramétrique. C'est pourquoi nous regroupons ces deux types de tests dans un même chapitre.

Une variable catégorielle est une extension d'une variable de Bernoulli au sens où il n'y a plus deux mais $c \geq 2$ résultats possibles. Il s'agit d'une variable aléatoire non pas à valeurs dans $\mathbb{R}$ comme les v.a. usuelles mais à valeurs dans un ensemble de catégories. Ce sera, par exemple, la réponse d'un individu à une question à c modalités dans une enquête par sondage. Une variable catégorielle peut être une variable purement qualitative (ou nominale) ou une variable ordinale si les catégories sont ordonnées. Elle peut aussi résulter d'une mise en catégories d'une variable quantitative (par exemple constitution de classes d'âge ou de revenus).

Une variable catégorielle est parfaitement définie par les probabilités respectives $p_1, p_2, \cdots, p_c$ des c catégories. La somme des probabilités étant égale à 1, il y a en vérité $c-1$ paramètres libres. Dans le contexte d'un tirage aléatoire d'un individu dans une population finie, exposé en section 2.4, $p_1, p_2, \cdots, p_c$ coïncident avec les fréquences relatives (ou proportions) des c catégories dans cette population.

D'une façon générale nous nous intéresserons à l'observation de la variable catégorielle sur un $n-$échantillon aléatoire, par exemple les catégories observées sur n individus tirés au hasard dans une population. Seules importent (comme

on le justifiera dans la note 10.1) les variables aléatoires $N_1, N_2, ..., N_c$ correspondant aux fréquences respectives des c catégories après n observations répétées et indépendantes de la variable catégorielle. Leur loi conjointe est la **loi multinomiale** décrite en section 4.1.6 en tant qu'extension de la loi binomiale de deux à c catégories. Cette loi sera donc le point de départ des tests concernant les variables catégorielles.

10.1 Test sur les paramètres d'une loi multinomiale

Dans les mêmes notations que ci-dessus, la fonction de probabilité conjointe des v.a. $N_1, N_2, \cdots, N_c$ est :

$$P(N_1 = n_1, N_2 = n_2, \cdots, N_c = n_c) = \frac{n!}{n_1! n_2! \cdots n_c!} p_1^{n_1} p_2^{n_2} \cdots p_c^{n_c}$$

si $\sum_{j=1}^{c} n_j = n$ et 0 sinon (voir section 4.1.6). Ces variables aléatoires ne sont pas indépendantes puisque leur somme doit être égale à n.

On s'intéresse à l'hypothèse nulle :

$$H_0 : p_1 = p_{01},\ p_2 = p_{02}, \cdots,\ p_c = p_{0c}$$

où les p_{0j} sont des valeurs de probabilités spécifiées telles que $\sum_{j=1}^{c} p_{0j} = 1$, l'hypothèse alternative étant qu'il existe au moins une catégorie j telle que $p_j \neq p_{0j}$ (en fait il y en aura au moins deux puisque le total doit rester égal à 1). Nous présentons deux approches de test dont on verra qu'elles sont asymptotiquement équivalentes.

10.1.1 Test du rapport de vraisemblance généralisé

Soit le RVG :

$$\lambda(n_1, n_2, \cdots, n_c) = \frac{\dfrac{n!}{n_1! n_2! \cdots n_c!}\ (p_{01})^{n_1} (p_{02})^{n_2} \cdots (p_{0c})^{n_c}}{\dfrac{n!}{n_1! n_2! \cdots n_c!}\ \widehat{p}_1^{\,n_1} \widehat{p}_2^{\,n_2} \cdots \widehat{p}_c^{\,n_c}}$$

où $(\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_c)$ est l'estimation du MV de $(p_1, p_2, \cdots, p_c)$. Montrons que $\widehat{p}_j$ est égal à n_j/n, la proportion observée dans l'échantillon pour la catégorie j. En remplaçant p_c par $1 - p_1 - \cdots - p_{c-1}$ pour intégrer la contrainte $\sum_{j=1}^{c} p_j = 1$ dans la recherche du maximum, la fonction de vraisemblance s'écrit :

$$\frac{n!}{n_1! n_2! \cdots n_c!} p_1^{n_1} p_2^{n_2} \cdots p_{c-1}^{n_{c-1}} (1 - p_1 - \cdots - p_{c-1})^{n_c} \quad \text{si } \sum_{j=1}^{c} n_j = n\,,$$

d'où la log-vraisemblance :

$$\ln(\frac{n!}{n_1!n_2!\cdots n_c!}) + n_1 \ln p_1 + \cdots + n_{c-1} \ln p_{c-1} + n_c \ln(1 - p_1 - \cdots - p_{c-1})$$

si $\sum_{j=1}^{c} n_j = n$.

En annulant la dérivée par rapport à chacun des paramètres $p_1, p_2, \cdots, p_{c-1}$ on obtient les $c-1$ équations suivantes :

$$\begin{cases} \dfrac{n_1}{p_1} - \dfrac{n_c}{1 - p_1 - \cdots - p_{c-1}} = 0 \\ \dfrac{n_2}{p_2} - \dfrac{n_c}{1 - p_1 - \cdots - p_{c-1}} = 0 \\ \qquad\qquad \cdots \\ \dfrac{n_{c-1}}{p_{c-1}} - \dfrac{n_c}{1 - p_1 - \cdots - p_{c-1}} = 0 \end{cases}$$

$$\text{soit} \qquad \frac{n_1}{p_1} = \frac{n_2}{p_2} = \frac{n_{c-1}}{p_{c-1}} = \frac{n_c}{p_c} = \frac{\sum_{j=1}^{c} n_j}{\sum_{j=1}^{c} p_j} = n\,,$$

d'où (en admettant que la solution unique aux équations donne bien un maximum), pour tout j, la solution $\widehat{p}_j = n_j/n$. □

Le rapport de vraisemblance pour tester H_0 est donc :

$$\begin{aligned} \lambda(n_1, n_2, \cdots, n_c) &= \frac{(p_{01})^{n_1}(p_{02})^{n_2}\cdots(p_{0c})^{n_c}}{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}\cdots\left(\frac{n_c}{n}\right)^{n_c}} \\ &= \prod_{j=1}^{c} \left(\frac{np_{0j}}{n_j}\right)^{n_j}. \end{aligned}$$

En utilisant le théorème asymptotique 9.2 on a comme région de rejet de niveau (approximativement) α :

$$-2\ln\lambda = 2\sum_{j=1}^{c} n_j \ln\frac{n_j}{np_{0j}} > \chi^{2\ (c-1)}_{1-\alpha}.$$

Notons que la loi du khi-deux a $c-1$ degrés de liberté, car il n'y a que $c-1$ paramètres libres spécifiés par H_0. Dans ce contexte de variables catégorielles la statistique $-2\ln\lambda$ est appelée *déviance.* Certains logiciels l'appellent toutefois rapport de vraisemblance (alors que celui-ci est λ). Par commodité nous nous autoriserons aussi ce glissement de langage.

Note 10.1 Chaque observation d'une variable catégorielle peut être décrite par un vecteur «indicateur» $(X_1, X_2, \cdots, X_c)$ tel que la j-ème composante prenne la valeur 1 si le résultat est la j-ème catégorie, les autres composantes prenant alors la valeur 0. La densité conjointe des c composantes est :

$$f(x_1, x_2, \cdots, x_c; p_1, p_2, \cdots, p_c) = p_1^{x_1} p_2^{x_2} \cdots p_c^{x_c}$$

où, pour tout j, $x_j \in \{0, 1\}$ et $\sum_{j=1}^{c} x_j = 1$. Pour n observations répétées on a une suite de n vecteurs : $\{(X_{1i}, X_{2i}, \cdots, X_{ci}),\ i = 1, \cdots, n\}$, avec densité conjointe :

$$\prod_{i=1}^{n} p_1^{x_{1i}} p_2^{x_{2i}} \cdots p_c^{x_{ci}} = p_1^{n_1} p_2^{n_2} \cdots p_c^{n_c}$$

si $\sum_{j=1}^{c} n_j = n$ (0 sinon) où n_j est le nombre d'observations tombant dans la catégorie j parmi les n observations. Donc $(N_1, N_2, \cdots, N_c)$ est une statistique exhaustive. De plus on voit que lorsque l'on fait un rapport de vraisemblance pour la suite des n vecteurs, on obtient la même expression qu'avec la loi multinomiale, les termes avec factoriels de cette dernière s'éliminant. Ceci justifie le fait de passer directement par la loi multinomiale pour construire le test.

10.1.2 Test du khi-deux de Pearson

Ce test est historiquement le premier à avoir été proposé bien avant le développement formel de la théorie des tests par Jerzy Neyman et par Egon Pearson à partir de 1930. Il a été mis au point vers 1900 par Karl Pearson, le père d'Egon, afin de vérifier sur des données biologiques certaines hypothèses tenant aux facteurs d'hérédité.

En utilisant des approximations gaussiennes Karl Pearson a montré que la statistique de test

$$Q = \sum_{j=1}^{c} \frac{(N_j - np_{0j})^2}{np_{0j}}$$

admet, sous H_0, une loi asymptotique $\chi^2(c-1)$. C'est pourquoi Q est couramment appelée *statistique du khi-deux (*ou *statistique de Pearson).*

Remarquons que, sous l'hypothèse H_0, $(N_j - np_{0j})/\sqrt{np_{0j}(1-p_{0j})}$ est la variable centrée-réduite de N_j dont la loi marginale est la loi binomiale $\mathcal{B}(n, p_{0j})$ (voir section 4.1.6). Asymptotiquement cette v.a. suit une loi $\mathcal{N}(0\,;1)$ et son carré une loi $\chi^2(1)$. On montre que la contrainte $\sum_{j=1}^{c} N_j = n$ a pour effet d'éliminer les facteurs $(1 - p_j^0)$ pour donner une loi asymptotique du khi-deux à $c-1$ degrés de liberté. Intuitivement on voit que la valeur prise par Q est d'autant plus petite que les fréquences observées sont proches des np_{0j} appelées *fréquences attendues* (ou fréquences théoriques) sous H_0. On ne rejette donc l'hypothèse nulle que pour de grandes valeurs de réalisation q de Q, à savoir lorsque $q > \chi^{2\ (c-1)}_{1-\alpha}$ pour un test de niveau (approximatif) α.

10.1.3 Équivalence asymptotique des deux tests

Nous donnons une démonstration abrégée de cette équivalence, des développements plus rigoureux se trouvant dans les ouvrages cités en référence (en fin de section 10.3). Pour ne pas alourdir les notations nous n'indiçons pas par n les suites de v.a. ou de réalisations dont on considère ici la convergence quand $n \to \infty$.

Pour toute composante N_j du vecteur aléatoire $(N_1, N_2, ..., N_c)$, sous H_0, $\frac{N_j}{n}$ tend presque sûrement vers p_{0j} quand $n \to \infty$ par la loi des grands nombres et, par conséquent, $\frac{N_j - np_{0j}}{np_{0j}} \xrightarrow{p.s.} 0$. Pour toute réalisation[1] $(n_1, n_2, ..., n_c)$ de $(N_1, N_2, ..., N_c)$ on a donc :

$$\frac{n_j - np_{0j}}{np_{0j}} \longrightarrow 0 \quad \text{quand } n \to \infty \text{ , pour tout } j.$$

Posant $h_j = \frac{n_j - np_{0j}}{np_{0j}}$ on peut écrire $n_j = np_{0j}(1 + h_j)$ et $\ln \frac{n_j}{np_{0j}} = \ln(1 + h_j)$. Pour chaque terme $n_j \ln \frac{n_j}{np_{0j}}$ dans l'expression de la réalisation de la déviance (i.e. $-2 \ln \lambda$), développons ce logarithme au voisinage de 1 selon $\ln(1 + x) = x - \frac{1}{2}x^2 + O(x^3)$ pour obtenir :

$$\begin{aligned} n_j \ln \frac{n_j}{np_{0j}} &= np_{0j}(1 + h_j)(h_j - \frac{1}{2}h_j^2 + O(h_j^3)) \\ &= np_{0j}\,(h_j + \frac{1}{2}h_j^2 + O(h_j^3)) \\ &= n_j - np_{0j} + \frac{1}{2}\frac{(n_j - np_{0j})^2}{np_{0j}} + np_{0j}O(h_j^3)\,. \end{aligned}$$

Le terme $np_{0j}O(h_j^3)$ qui est négligeable devant les deux autres termes est d'ordre :

$$(np_{0j})\frac{(n_j - np_{0j})^3}{(np_{0j})^3} = \left[\frac{n_j - np_{0j}}{\sqrt{np_{0j}(1 - p_{0j})}}\right]^3 \frac{(1 - p_{0j})^{3/2}}{\sqrt{np_{0j}}}.$$

L'expression entre crochets étant une réalisation d'une variable aléatoire $\mathcal{N}(0\,;1)$, le terme négligé est d'ordre $1/\sqrt{n}$.

La réalisation de la déviance est donc :

$$2\sum_{j=1}^{c} n_j \ln \frac{n_j}{np_{0j}} = 2\sum_{j=1}^{c}(n_j - np_{0j}) + \sum_{j=1}^{c}\frac{(n_j - np_{0j})^2}{np_{0j}} + O(\frac{1}{\sqrt{n}}).$$

Comme $\sum_{j=1}^{c}(n_j - np_{0j}) = 0$ on voit que la réalisation de la déviance est équivalente à la réalisation q de la statistique du khi-deux quand $n \to \infty$. Ceci

[1] Plus formellement, se référant à la convergence presque sûre, on devrait dire «pour presque toute réalisation».

a pour conséquence que **les statistiques du rapport de vraisemblance**[2] **et du khi-deux ont la même loi asymptotique** sous H_0. Elles conduisent à des régions de rejet identiques et, donc, à des tests équivalents.

Note 10.2 Il est important de remarquer que le seul point crucial de la démonstration de l'équivalence asymptotique des deux statistiques est que $(n_j - np_{0j})/np_{0j}$ tende vers 0 quand $n \to \infty$ ou, de même, que $(n_i/n)/p_{0j}$ tende vers 1. Ainsi on pourrait avoir en lieu et place de p_{0j} une estimation convergente de cette valeur. Dans les sections suivantes on aura à effectuer de telles estimations et on admettra alors que les deux statistiques conservent la même loi asymptotique.

En pratique on préfère utiliser le test du khi-deux qui met en évidence les écarts entre les fréquences observées et les fréquences attendues. D'autant plus que la statistique du khi-deux converge plus vite que celle du RV et donne donc une meilleure approximation (voir Agresti, 2002). On considère généralement, comme conditions de validité de l'approximation asymptotique, que **les fréquences attendues np_{0j} doivent rester supérieures à 5**. Lorsque ces conditions ne sont pas remplies on s'arrange pour regrouper certaines catégories proches.

Nous ne nous intéresserons pas à la puissance de ce test qui est un problème complexe vu la multiplicité de formes que peut revêtir l'hypothèse alternative. Ceci sera vrai a fortiori pour les tests introduits dans les sections suivantes.

10.1.4 Cas particulier de la loi binomiale

Appliquons la formule du khi-deux avec $c = 2$, $p_{01} = p_0$, $p_{02} = 1 - p_0$ et $n_2 = n - n_1$. On a :

$$\begin{aligned} q &= \frac{(n_1 - np_0)^2}{np_0} + \frac{(n - n_1 - n(1-p_0))^2}{n(1-p_0)} \\ &= \frac{(n_1 - np_0)^2}{np_0} + \frac{(n_1 - np_0)^2}{n(1-p_0)} \\ &= \frac{(n_1 - np_0)^2}{np_0(1-np_0)} = \frac{(\widehat{p} - p_0)^2}{\dfrac{p_0(1-p_0)}{n}} \end{aligned}$$

en posant $\widehat{p} = n_1/n$ pour la fréquence relative de succès observée.

Remarquons maintenant que si Z est une v.a. de loi $\mathcal{N}(0\,;1)$ alors, avec les notations usuelles pour les quantiles de cette loi :

$$P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = P(Z^2 < (z_{1-\alpha/2})^2) = 1 - \alpha.$$

[2]Plus exactement la statistique de la déviance.

Comme Z^2 suit une loi $\chi^2(1)$ on a, pour les quantiles, l'égalité $(z_{1-\alpha/2})^2 = \chi^{2\,(1)}_{1-\alpha}$. Ainsi la région d'acceptation $q < \chi^{2\,(1)}_{1-\alpha}$ du test du khi-deux est identique à :

$$-z_{1-\alpha/2} < \frac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} < z_{1-\alpha/2}$$

qui est celle du test classique par approximation gaussienne pour une proportion proposé en section 9.7.5 pour le test bilatéral $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$.

10.2 Test de comparaison de plusieurs lois multinomiales

Ce test est une double généralisation du test de comparaison de deux lois de Bernoulli vu en section 9.7.6, considérant à la fois plusieurs catégories et plusieurs lois.

Soit J lois multinomiales ayant les mêmes catégories, en nombre I, et soit p_{ij} la probabilité d'être dans la catégorie i pour la loi j. L'hypothèse nulle à tester est que les probabilités associées aux I catégories sont identiques pour toutes les J lois, soit :

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{iJ} \quad , \quad i = 1, \ldots, I \ ,$$

l'hypothèse alternative étant que pour au moins une catégorie (en fait il y en aura au moins deux puisque le total doit rester égal à 1) ces probabilités diffèrent pour au moins deux lois. Si l'on se réfère à la comparaison de populations ce test est un test d'**homogénéité des populations** au sens où H_0 signifie que la variable catégorielle étudiée se distribue de façon identique dans ces populations (voir les exercices appliqués pour illustration).

Dans ce problème on est en présence de $(I-1)J$ paramètres inconnus dont seulement $(I-1)(J-1)$ sont spécifiés par H_0. En effet, pour la catégorie i par exemple, on peut écrire, en prenant la J-ème loi pour référence, $p_{i1} = p_{iJ} + \theta_1$, $p_{i2} = p_{iJ} + \theta_2, \ldots, p_{i,J-1} = p_{iJ} + \theta_{J-1}$. Pour cette catégorie, H_0 équivaut donc à $\theta_1 = \theta_2 = \ldots = \theta_{J-1} = 0$. Comme il suffit que ce type d'égalité soit vérifié pour $I-1$ catégories, l'égalité étant nécessairement vérifiée pour la catégorie restante, on a bien $(I-1)(J-1)$ paramètres spécifiés par H_0.

On considère J échantillons mutuellement indépendants de tailles n_j où $j = 1, 2, \ldots, J$, issus respectivement des J lois. Soit N_{ij} la fréquence de la catégorie i pour la loi multinomiale j, $N_{i.}$ le total des fréquences pour la catégorie i sur l'ensemble des lois et $n = \sum_{j=1}^{J} n_j$ l'effectif englobant tous les J échantillons. On pourrait développer aisément le test du rapport de vraisemblance. Par application du théorème 9.2, la statistique du RVG suit asymptotiquement, sous H_0, une loi du khi-deux à $(I-1)(J-1)$ degrés de liberté. Comme

précédemment nous préférons utiliser la statistique Q de Pearson. Toutefois, ici, les p_{ij} ne sont pas totalement spécifiés sous H_0 et il faut les estimer pour calculer les fréquences attendues. En recourant aux estimateurs du maximum de vraisemblance qui sont convergents, les deux tests restent asymptotiquement équivalents (voir note 10.2). La loi asymptotique de Q sera donc la loi du khi-deux à $(I-1)(J-1)$ degrés de liberté.

Sous H_0 notons p_i la probabilité de la catégorie i, commune à toutes les lois (soit $p_i = p_{i1} = p_{i2} = \cdots = p_{iJ}$) et déterminons les estimateurs du MV des p_i. Pour un échantillon, disons l'échantillon j, la vraisemblance est, comme vu au début de la section 10.1.1,

$$\frac{n_j!}{n_{1j}!n_{2j}!\cdots n_{Ij}!} p_1^{n_{1j}} p_2^{n_{2j}} \cdots p_{I-1}^{n_{I-1,j}} (1-p_1-\cdots-p_{I-1})^{n_{Ij}} \quad \text{si} \sum_{i=1}^{J} n_{ij} = n_j \,.$$

Pour l'ensemble des J échantillons la vraisemblance globale est le produit des vraisemblances de chaque échantillon puisque ceux-ci sont indépendants, soit :

$$\begin{aligned} &\prod_{j=1}^{J} \frac{n_j!}{n_{1j}!n_{2j}!\cdots n_{Ij}!} p_1^{n_{1j}} p_2^{n_{2j}} \cdots p_{I-1}^{n_{I-1,j}} (1-p_1-\cdots-p_{I-1})^{n_{Ij}} \\ &= \left(\prod_{j=1}^{J} \frac{n_j!}{n_{1j}!n_{2j}!\cdots n_{Ij}!} \right) p_1^{n_{1\cdot}} p_2^{n_{2\cdot}} \cdots p_{I-1}^{n_{I-1,\cdot}} (1-p_1-\cdots-p_{I-1})^{n_{I\cdot}} \,, \end{aligned}$$

où $n_{i\cdot}$ est le total observé pour la catégorie i sur toutes les lois.

On est ramené au même problème de maximisation qu'en section 10.1.1, les $n_{i\cdot}$ se substituant aux n_i . Le maximum est donc atteint pour $\widehat{p}_i = n_{i\cdot}/n$. L'estimateur du maximum de vraisemblance de p_i est donc l'estimateur naturel $N_{i\cdot}/n$ égal à la fréquence relative de la catégorie i obtenue en fusionnant les J échantillons. Pour la catégorie i de la loi j la fréquence attendue sous H_0 est donc estimée par $n_j N_{i\cdot}/n$. La statistique de test est alors :

$$Q = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(N_{ij} - \dfrac{N_{i\cdot} n_j}{n})^2}{\dfrac{N_{i\cdot} n_j}{n}}$$

dont la loi peut être approchée par la loi $\chi^2((I-1)(J-1))$.

Pour la mise en oeuvre de ce test par la statistique du khi-deux considérons les notations pour les fréquences observées comme indiqué dans le tableau suivant :

Catégorie	loi 1	$\cdots$	loi j	$\cdots$	loi J	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

Dans ce tableau les n_j sont notés $n_{.j}$ pour donner un rôle symétrique aux deux marges. La fréquence attendue pour la case (i, j) est obtenue en effectuant le produit des marges $n_{i.}$ et $n_{.j}$ divisé par n. On rejettera donc H_0 au niveau α si :

$$q = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} > \chi^2_{1-\alpha}{}^{((I-1)(J-1))} .$$

Pour un tableau 2×2 la condition de validité reste que les fréquences attendues soient supérieures à 5. Pour un tableau de dimensions supérieures de nombreuses simulations ont montré que l'approximation était étonnamment bonne même avec des effectifs plus faibles. On montre aisément (voir exercices) que, dans le cas $I = 2$ et $J = 2$, on retombe sur le test par approximations gaussiennes de la section 9.7.6.

10.3 Test d'indépendance de deux variables catégorielles

10.3.1 Test du RVG et test du khi-deux

On considère un couple de variables catégorielles, l'une à I catégories, l'autre à J catégories, observables sur toute unité statistique sélectionnée (ou, pour un sondage, sur chaque individu d'une population). Le croisement de ces deux variables donne lieu à une variable catégorielle à $I \times J$ catégories avec $I \times J - 1$ paramètres libres. A la catégorie obtenue par croisement des catégories i et j respectives de chaque variable est associée la probabilité p_{ij}. On a donc $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$. On s'intéresse à l'hypothèse d'indépendance de ces deux variables.

Comme pour les variables aléatoires (voir définition 3.4) on définit que deux variables catégorielles sont indépendantes par le fait que, pour tout événement sur l'une et tout événement sur l'autre, la probabilité de leur intersection (ou conjonction) est égale au produit des probabilités de chaque événement. Un événement étant un sous-ensemble de catégories et les catégories étant en nombre

fini on peut voir aisément (voir exercices) qu'il faut et qu'il suffit qu'il y ait indépendance entre tous les couples élémentaires (i,j) de catégories pour assurer l'indépendance complète. Ainsi l'hypothèse nulle à tester est :

$$H_0 : p_{ij} = p_{i.}p_{.j} \quad \text{pour } i = 1, \cdots, I \quad \text{et} \quad j = 1, \cdots, J\,,$$

où $p_{i.}$ est la probabilité marginale pour la catégorie i de la première variable et $p_{.j}$ pour la catégorie j de la deuxième variable. L'hypothèse alternative est la négation de H_0 à savoir qu'il y ait au moins un couple (i,j) pour lequel $p_{ij} \neq p_{i.}p_{.j}$.

Pour un échantillon aléatoire de taille n on observe les fréquences au croisement des deux variables et l'on note N_{ij} la fréquence au croisement (i,j). Une réalisation de l'échantillon peut être représentée par le *tableau de contingence* suivant :

Var.1 \ Var.2	1	$\cdots$	j	$\cdots$	J	
1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	n_{I1}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I.}$
	$n_{.1}$	$\cdots$	$n_{.j}$	$\cdots$	$n_{.J}$	n

Prenons l'approche par le test du rapport de vraisemblance. Considérant la variable catégorielle à $I \times J$ catégories obtenue par le croisement on a pour fonction de vraisemblance du tableau des n_{ij} (voir section 10.1) :

$$\frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} p_{ij}^{n_{ij}} \quad \text{pour } \sum_{i,j} n_{ij} = n \quad \text{et} \quad 0 \text{ sinon},$$

où $\prod_{i,j}$ dénote, en abrégé, les produits de $I \times J$ termes avec $i = 1, \cdots, I$ et $j = 1, \cdots, J$ (et de même $\sum_{i,j}$ pour les sommes). Comme il a été démontré en section 10.1.1 les estimations du maximum de vraisemblance sont $\widehat{p}_{ij} = n_{ij}/n$.

Sous H_0 la vraisemblance est :

$$\begin{aligned} \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} (p_{i.}p_{.j})^{n_{ij}} &= \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} {p_{i.}}^{n_{ij}} \prod_{i,j} {p_{.j}}^{n_{ij}} \\ &= \frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i} {p_{i.}}^{n_{i.}} \prod_{i} {p_{.j}}^{n_{.j}} . \end{aligned}$$

La maximisation s'opère séparément sur chaque variable et revient pour chacune au problème d'une multinomiale d'où $\widehat{p}_{i.} = n_{i.}/n$ et $\widehat{p}_{.j} = n_{.j}/n$. Le RVG est donc :

$$\lambda = \frac{\prod_i {n_{i.}}^{n_{i.}} \prod_j {n_{.j}}^{n_{.j}}}{n^n \prod_{i,j} n_{ij}^{n_{ij}}}.$$

Asymptotiquement la statistique $-2\ln\lambda$ (obtenue en remplaçant les réalisations n_{ij} par les v.a. N_{ij} dans l'expression ci-dessus) suit une loi du khi-deux avec un nombre de degrés de liberté égal au nombre de paramètres spécifiés. On pourrait trouver ce nombre en utilisant une reparamétrisation comme nous l'avons fait en section 10.2, mais il est plus simple de constater que sous H_0 il reste $I-1$ paramètres inconnus pour la première variable et $J-1$ pour la deuxième. Comme il y a globalement $IJ-1$ paramètres inconnus cela signifie que H_0 spécifie implicitement $IJ-1-(I-1)-(J-1) = (I-1)(J-1)$ paramètres. Donc la statistique $-2\ln\lambda$ suit, sous H_0, une loi $\chi^2((I-1)(J-1))$ et l'on rejettera H_0 au niveau α si la réalisation $-2\ln\lambda$ est supérieure à $\chi^{2\,((I-1)(J-1))}_{1-\alpha}$.

De préférence, on utilisera la statistique Q de Pearson obtenue en estimant les fréquences attendues sous H_0 par le maximum de vraisemblance, soit :

$$n\widehat{p}_{i.}\widehat{p}_{.j} = n\frac{N_{i.}}{n}\frac{N_{.j}}{n} = \frac{N_{i.}N_{.j}}{n},$$

d'où :

$$Q = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{(N_{ij} - \frac{N_{i.}N_{.j}}{n})^2}{\frac{N_{i.}N_{.j}}{n}}.$$

Cette statistique est asymptotiquement de même loi que celle issue du RVG (voir note 10.2) sous H_0. Cette hypothèse sera donc rejetée au niveau α si la réalisation q de Q est telle que $q > \chi^{2\,((I-1)(J-1))}_{1-\alpha}$.

Ceci conduit à **un test dont la mise en oeuvre est en tous points identique à celui proposé pour la comparaison de lois multinomiales.** Il convient toutefois d'insister sur la différence entre les deux situations. Dans le test d'indépendance seule est fixée la taille globale de l'échantillon n, les fréquences marginales étant aléatoires. Dans la situation précédente une des marges (celle du bas du tableau) est fixée dans le plan d'échantillonnage par les effectifs choisis pour les différentes populations.

Notons aussi que le tableau des fréquences attendues (les $n_{i.}n_{.j}/n$) est un tableau dont toutes les lignes (respectivement toutes les colonnes) sont proportionnelles, ce qui correspond bien à l'idée d'indépendance intuitive sur un tableau empirique.

10.3.2 Test exact de Fisher (tableau 2×2)

On considère le cas $I = 2$ et $J = 2$ avec le tableau des réalisations suivant :

Var.1 \ Var.2	1	2	
1	n_{11}	n_{12}	$n_{1.}$
2	n_{21}	n_{22}	$n_{2.}$
	$n_{.1}$	$n_{.2}$	n

On peut montrer que la loi conditionnelle conjointe de $(N_{11}, N_{12,}N_{21,}N_{22})$ sachant les fréquences marginales $N_{1.}, N_{2.}, N_{.1}, N_{.2}$ est indépendante des p_{ij} sous H_0. En d'autres termes les fréquences marginales sont des statistiques exhaustives pour les paramètres $(p_{1.}, p_{2.}, p_{.1}, p_{.2})$ en cas d'indépendance (voir définition 6.8). Plus précisément, en raison des contraintes, $(N_{1.}, N_{.1})$ est statistique exhaustive si l'on prend comme seuls paramètres inconnus $(p_{1.}, p_{.1})$. La démonstration est simple car, les marges étant fixées à $n_{1.}, n_{2.}, n_{.1}, n_{.2}$, il suffit de considérer la probabilité $P(N_{11} = n_{11}|n_{1.}, n_{2.}, n_{.1}, n_{.2})$ puisque les v.a. $N_{12,}N_{21,}N_{22}$ sont liées à N_{11} respectivement par $n_{1.} - N_{11}, n_{.1} - N_{11}$ et $n_{2.} - n_{1.} + N_{11}$. On a alors une démonstration analogue à celle exposée en section 9.7.6 (voir exercices). On obtient :

$$P(N_{11} = n_{11}|n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{12}}}{\binom{n}{n_{1.}}}$$

qui montre que, conditionnellement aux marges, N_{11} suit une loi hypergéométrique $\mathcal{H}(n, n_{1.,}n_{.1})$.

Le test proposé par Fisher consiste à prendre une région critique de niveau α choisi, sur cette **loi conditionnelle**. On peut établir (voir note 10.3 ci-dessous) que ceci est légitime, donnant bien un test de niveau α dans l'absolu. Cela conduit à une règle de décision totalement identique à celle utilisée en section 9.7.6 pour tester de façon exacte l'égalité de deux proportions (ou, plus généralement, des paramètres de deux lois de Bernoulli). Pour cette hypothèse-là il suffisait toutefois de conditionner sur une seule marge, l'autre étant fixée par le plan d'échantillonnage.

On définit donc une région de rejet sur la base, par exemple, de la valeur n_{11} observée, selon :

$$n_{11} \notin [c_{\alpha_1}, c_{\alpha_2}]$$

où c_{α_1} et c_{α_2} sont les quantiles d'ordres α_1 et α_2 tels que $\alpha_1 + \alpha_2 = \alpha$, issus de la loi $\mathcal{H}(n, n_{1.,}n_{.1})$. Ce test est UPP-sans biais.

La plupart des logiciels se chargent d'effectuer les calculs fastidieux et fournissent la P-valeur relative au tableau observé.

On pourrait étendre ce test exact à un tableau de plus grande dimension. Outre que les calculs se complexifient rapidement ceci n'est pas utile du fait que l'approximation du test du khi-deux devient très vite satisfaisante avec des conditions de validité identiques à celles de la section précédente, à savoir que les fréquences attendues restent pour la plupart supérieures ou égales à 5.

Note 10.3 Montrons que, d'une façon générale, un test de niveau α conditionnellement à une statistique exhaustive sous H_0 est un test de niveau α dans l'absolu. Soit dans la famille $\{f(x;\theta),\, \theta \in \Theta\}$ l'hypothèse nulle $H_0 : \theta \in \Theta_0$ et T une statistique exhaustive sous H_0. Soit, pour un échantillon $(X_1, X_2, \cdots, X_n)$, un test défini par une région de rejet $\overline{A}$ de $\mathbb{R}^n$ de niveau α pour la loi conditionnelle $f(x_1, x_2, \cdots, x_n | T = t)$ qui, par définition pour T, ne dépend pas de θ sous H_0. Alors, notant en raccourci $(x_1, x_2, \cdots, x_n)$ par $\mathbf{x}$ et $dx_1 dx_2 \cdots dx_n$ par $d\mathbf{x}$, on a pour tout $\theta \in \Theta_0$, avec des notations évidentes (T étant de dimension k) :

$$\begin{aligned} P_\theta(\overline{A}) &= \int_{\overline{A}} f(\mathbf{x};\theta) d\mathbf{x} \\ &= \int_{\overline{A}} \left[\int_{\mathbb{R}^k} f(\mathbf{x}|\, T = t) f_T(t;\theta) dt \right] d\mathbf{x} \\ &= \int_{\mathbb{R}^k} f_T(t;\theta) \left[\int_{\overline{A}} f(\mathbf{x}|\, T = t) d\mathbf{x} \right] dt \\ &= \alpha \int_{\mathbb{R}^k} f_T(t;\theta) dt = \alpha \ , \end{aligned}$$

ce qui prouve que $\overline{A}$ définit un test de niveau α de façon non conditionnelle. Dans le cas discret ceci vaut si l'on randomise le test pour atteindre exactement le niveau α. Si l'on utilise un test conservateur conditionnellement il restera toutefois conservateur non conditionnellement.

Exemple 10.1 Une pré-enquête a été effectuée auprès de 50 personnes (supposées sélectionnées par sondage aléatoire simple dans la population cible) pour évaluer le taux d'acceptation pour participer à une étude de suivi médical. On s'intéresse au croisement des variables catégorielles sexe et réponse oui/non pour participer. On a obtenu les résultats suivants :

sexe \ participation	oui	non	
femme	9	20	29
homme	5	16	21
	14	36	50

Choisissant de fonder le test sur la fréquence du croisement (femme, oui) on doit lire les probabilités sur une loi $\mathcal{H}(50, 29, 14)$. C'est-à-dire qu'on examine la

v.a. X «nombre de femmes parmi les 14 oui» sachant qu'il y a globalement 29 femmes parmi les 50 personnes, sous l'hypothèse d'indépendance entre sexe et participation ou non. On doit rejeter cette hypothèse si le nombre de femmes répondant oui est soit trop élevé, soit trop faible, par rapport à la fréquence attendue qui est égale à $14\frac{29}{50} = 8{,}12$. Pour cette loi hypergéométrique les sauts de probabilité sont assez élévés. On a, par exemple, une région de rejet de niveau 0,024 avec $\{0, 1, 2, 3, 4,\ 12, 13, 14\}$ pour les valeurs de X et de niveau 0,11 avec $\{0, 1, 2, 3, 4, 5, 11, 12, 13, 14\}$. On a donc avantage à considérer la P-valeur. Avec la fonction « loi hypergéométrique » dans un tableur on trouve que $P(X \geq 9) = 0{,}41$. La P-valeur est donc égale à 0,82 ce qui rend l'hypothèse d'indépendance tout à fait acceptable.

Notons qu'on aurait pu aussi bien prendre la loi $\mathcal{H}(50, 14, 29)$, considérant le nombre de oui parmi les 29 femmes sachant qu'il y a globalement 14 oui parmi les 50 personnes, ce qui est rigoureusement équivalent.

Voyons ce que donne l'approche approximative par la statistique de Pearson. Le tableau des fréquences attendues est

8,12	20,88
5,88	15,12

d'où :

$$q = (0{,}88)^2\left(\frac{1}{8{,}12} + \frac{1}{20{,}88} + \frac{1}{5{,}88} + \frac{1}{15{,}12}\right) = 0{,}315$$

qui correspond au quantile d'ordre 0,43 sur une loi $\chi^2(1)$. La P-valeur est donc ici donnée à 0,57 ce qui est sensiblement différent de la valeur exacte de 0,82 mais conduit à la même décision d'acceptation de l'indépendance.

Sur cet exemple, de la façon dont les choses sont présentées, on a le sentiment que l'on compare deux proportions : celle des femmes et celle des hommes acceptant de participer à l'étude. Néanmoins il ne s'agit pas d'un test d'égalité entre ces deux proportions car un tel test supposerait que l'on ait fixé a priori la taille des échantillons de chaque sexe (par un plan de *sondage stratifié* selon le sexe), alors que dans notre exemple les effectifs des hommes et des femmes résultent du tirage au hasard. La différence mérite d'être précisée même si elle n'a pas d'incidence sur la règle de décision. Elle n'est toutefois pas neutre pour le calcul de la puissance. ■

Pour approfondir la théorie et la pratique des données catégorielles on pourra consulter les ouvrages suivants : Agresti (2002), Chap (1998), Droesbeke, Lejeune et Saporta (2004).

10.4 Test d'ajustement à un modèle de loi

Le problème envisagé ici est de décider, au vu d'un échantillon $X_1, X_2, \ldots, X_n$, si la loi mère de cet échantillon est du type spécifié par une hypothèse

H_0. Un test aura pour but d'examiner si la distribution des valeurs de l'échantillon s'ajuste suffisamment bien à une distribution théorique donnée. On parle également de **test d'adéquation** (en anglais : *goodness-of-fit* test). A défaut de pouvoir rejeter H_0 on acceptera le modèle théorique proposé. Les développements précédents relatifs à des comparaisons de fréquences observées et de fréquences attendues ou «théoriques» vont nous fournir un test de portée très générale. Outre ce test fondé sur une statistique du khi-deux il existe bien d'autres tests d'ajustement d'inspirations diverses et nous étudierons en particulier le test de Kolmogorov-Smirnov de portée également générale et, par là-même, très répandu.

Nous distinguons deux situations pour l'hypothèse nulle. En premier nous étudions le cas plus simple où la loi est parfaitement spécifiée par H_0, puis nous passerons au cas où H_0 spécifie une famille paramétrique particulière sans préciser la valeur du paramètre qui reste inconnu.

10.4.1 Ajustement à une loi parfaitement spécifiée

Nous nous plaçons dans une optique **non paramétrique** au sens où les tests considérés devront s'appliquer quelle que soit la nature du modèle de loi mère envisagé.

La fonction de répartition étant l'objet mathématique le plus approprié pour spécifier une loi, qu'elle soit discrète ou continue, nous conviendrons d'écrire l'hypothèse nulle sous la forme $H_0 : F = F_0$ où F_0 caractérise donc le modèle de loi spécifié, l'alternative étant $H_1 : F \neq F_0$. Ce genre de situation n'est pas rare, par exemple lorsqu'une théorie a été élaborée pour un phénomène quantifiable et qu'il s'agit de la mettre à l'épreuve des faits.

Test du khi-deux

Son principe repose sur la transformation de la variable aléatoire en une variable catégorielle pour se ramener au test sur une loi multinomiale comme en section 10.1. Pour cela on découpe $\mathbb{R}$ (ou sa partie utile) en intervalles pour obtenir des classes comme on le ferait pour un histogramme. A l'instar de ce qui a été fait en section 8.5.2 ce découpage se définit comme une suite double de valeurs croissantes $\{\cdots, a_{-i}, \cdots, a_{-1}, a_0, a_1, \cdots, a_i, \cdots\}$ et l'on note n_k la fréquence des observations situées dans l'intervalle $]a_{k-1}, a_k]$ pour un échantillon de taille n. La fréquence attendue sous H_0 est np_k où p_k est la probabilité pour la loi F_0 associée à l'intervalle $]a_{k-1}, a_k]$, i.e.

$$p_k = F_0(a_k) - F_0(a_{k-1}).$$

Pour ce découpage il faut toutefois veiller à remplir les conditions de validité de l'approximation asymptotique, à savoir faire en sorte que les np_k restent supérieurs ou égaux à 5. Cela amènera à reconsidérer éventuellement

le découpage initial et à regrouper des classes contiguës à faible probabilité. Pour une extrémité infinie on constituera un dernier intervalle ouvert sur l'infini de probabilité supérieure à $5/n$. Notons que dans le cas d'une loi discrète concentrée sur un faible nombre de valeurs, chaque valeur (sauf peut-être sur les extrémités) peut constituer naturellement une classe en soi.

Par ailleurs le choix du nombre de classes, voire même des frontières de ces classes, influe sur la puissance. Mais il est difficile d'orienter ce choix et l'on recommande, pour la pratique, de rester proche de classes à probabilités égales.

Remarquons que le passage d'une variable aléatoire (au sens strict, c'est-à-dire quantitative) à une variable catégorielle induit une perte d'information. En effet dans une variable catégorielle il n'y a pas d'ordre des catégories et la statistique du khi-deux est indifférente à une permutation de celles-ci. Ainsi l'échelle numérique de la variable aléatoire est ignorée ce qui laisse supposer une perte de puissance.

Test de Kolmogorov-Smirnov

Ce test tient compte de l'échelle des observations mais ne s'applique en principe qu'aux **lois continues.** Il est fondé sur l'écart constaté entre la fonction de répartition empirique F_n et F_0. Nous avons vu en section 8.5.3 diverses propriétés de la fonction de répartition empirique comme estimateur de la fonction de répartition de la loi mère : elle est l'estimateur fonctionnel du maximum de vraisemblance et est convergente presque sûrement uniformément (voir théorème 8.1). De plus, $F_n(x)$ est sans biais en tout x fixé. On s'attend donc, si H_0 est vraie, à ce qu'elle reste proche de F_0. En fait le théorème 8.2 dû à Kolmogorov et à Smirnov fournit la statistique de test

$$D_n = \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| .$$

Son intérêt est que, sous H_0, sa loi ne dépend pas de la nature de F_0 ce qui donne lieu à des tables uniques quel que soit le type de modèle à tester.

Rappelons, suite au théorème 8.2, que pour $n \geq 40$ et $x >$ `0,8` on peut utiliser :

$$P(\sqrt{n}D_n < x) \simeq 1 - 2e^{-2x^2} ,$$

ce qui conduit, par exemple, à $P(D_n < \frac{1{,}36}{\sqrt{n}}) \simeq 0{,}95$. Comme on rejette naturellement H_0 si D_n est trop grand, on a comme région de rejet au niveau $0{,}05 : d_n > \frac{1{,}36}{\sqrt{n}}$ où d_n est la réalisation observée de D_n.

D'un point de vue pratique il faut tenir compte du fait que F_n est constante par morceaux. S'il est clair qu'en raison de la croissance de F_0 l'écart absolu maximal doit se situer en un point de discontinuité de F_n (donc en une valeur observée x_i) il y a toutefois lieu de comparer, pour tout $i = 1, \ldots, n$, la valeur de $F_0(x_i)$ à la fois à $F_n(x_i)$ et à $F_n(x_i^-) = F_n(x_{i-1})$. Une illustration est donnée en figure 10.1.

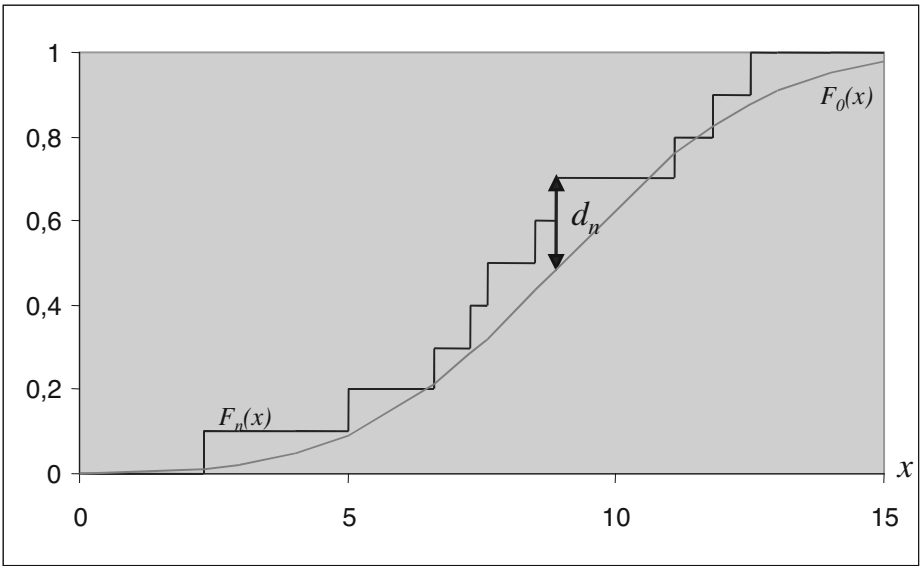


Figure 10.1 - Illustration du test de Kolmogorov-Smirnov.

On peut appliquer le test à des observations regroupées en classes. Dans ce cas il ne faut comparer que les valeurs aux frontières des classes : si a_k est une valeur frontière on comparera uniquement $F_n(a_k)$ et $F_0(a_k)$. En effet les valeurs ne sont comptabilisées que sur ces frontières et on ignore l'allure de F_n à l'intérieur des classes. Il se trouve qu'ainsi le test est conservateur (i.e. le niveau réel reste inférieur au niveau nominal de la table).

De nombreuses études ont été effectuées pour comparer les puissances du test de Kolmogorov-Smirnov et du test du khi-deux. Bien qu'on ne puisse tirer de conclusions générales il est vrai que le plus souvent le test du khi-deux est moins puissant. Ceci s'explique notamment par le fait que, contrairement au test de Kolmogorov-Smirnov, il ne tient pas compte de l'échelle des valeurs. Il est par ailleurs intéressant de noter que si les deux tests sont convergents, ils ne sont pas sans biais vis-à-vis de toutes les fonctions de répartitions autres que F_0 tant leur multiplicité est grande.

10.4.2 Ajustement dans une famille paramétrique donnée

On suppose maintenant, comme c'est le plus souvent le cas en pratique, que H_0 spécifie une famille de loi paramétrique sans précision sur le paramètre de la loi qui reste inconnu, ce que nous pouvons écrire :

$$H_0 : F \in \{f(x;\theta),\, \theta \in \Theta\}$$

où seul θ est inconnu. Par exemple on souhaite tester que la loi mère est gaussienne (ou, plus exactement, que le modèle gaussien est acceptable au vu des observations dont on dispose). Pour pouvoir élaborer une statistique mesurant d'une certaine façon, que ce soit par la statistique du khi-deux ou par celle de Kolmogorov-Smirnov, l'écart entre ce que l'on a observé et une référence théorique sous H_0, il est nécessaire de passer par une estimation du paramètre inconnu θ.

Test du khi-deux

Pour la statistique du khi-deux le théorème ci-après dont la démonstration a été effectuée par Cramer (1946) permet de prendre en compte cette estimation de θ. Donnons tout d'abord le cadre général d'application du théorème.

On considère une loi multinomiale à c catégories dont les probabilités **dépendent d'un paramètre inconnu** $\theta \in \Theta$ de dimension $r < c-1$ et sont notées $p_j(\theta)$, $j = 1, \cdots, c$. Pour un n-échantillon aléatoire et les fréquences observées $n_1, n_2, ..., n_c$, les estimations du MV de ces probabilités se déduisent, en tant que fonctions de θ, de l'estimation du maximum de vraisemblance $\widehat{\theta}$ de θ (voir note 6.4). Celui-ci est obtenu en maximisant la fonction de vraisemblance

$$L(\theta) = \frac{n!}{n_1!n_2!\cdots n_c!}[p_1(\theta)]^{n_1}[p_2(\theta)]^{n_2}\cdots[p_c(\theta)]^{n_c}$$

avec la contrainte $\sum_{j=1}^{c} p_j(\theta) = 1$. On en déduit alors les estimations du maximum de vraisemblance $p_1(\widehat{\theta}), p_2(\widehat{\theta}), \cdots, p_c(\widehat{\theta})$. Dans l'énoncé du théorème nous utilisons, pour alléger, ces mêmes notations pour les estimateurs, les n_i devant être remplacés par les N_j.

Théorème 10.1 *Soit une loi multinomiale à c catégories de probabilités $p_1(\theta), p_2(\theta), \cdots, p_c(\theta)$, où θ est un paramètre inconnu de dimension $r < c-1$ et soit $\widehat{\theta}$ l'estimateur du maximum de vraisemblance de θ pour les v.a. $N_1, N_2, ...,$ N_c. Alors (sous certaines conditions de régularité) on a :*

$$Q = \sum_{j=1}^{c} \frac{(N_j - np_j(\widehat{\theta}))^2}{np_j(\widehat{\theta})} \rightsquigarrow \chi^2(c-r-1).$$

En fait, nous avons déjà rencontré une telle situation dans le test d'indépendance en section 10.3. En effet sous l'hypothèse d'indépendance H_0 les p_{ij} de la variable catégorielle croisée à $I \times J$ catégories s'exprimaient selon $p_{ij} = p_{i.}p_{.j}$ et étaient donc des fonctions de $(I-1)+(J-1)$ paramètres correspondant aux probabilités marginales $p_{i.}$, $i = 1, \cdots, I-1$ et $p_{.j}$, $j = 1, \cdots, J-1$ ($p_{I.}$ et $p_{.J}$ se déduisant des précédents). On retrouve ici la règle des degrés de libertés pour Q, à savoir $IJ - (I-1) - (J-1) - 1 = (I-1)(J-1)$ établie alors en se rapportant à l'équivalence avec la statistique du RVG. On pourrait penser

que ce théorème est superflu, mais il mérite d'être présenté du fait qu'il découle historiquement d'une approche directe du comportement de la statistique du khi-deux et ceci, c'est à noter, sans faire référence à un test d'hypothèse.

Si l'on se replace dans une situation de test on peut au premier abord s'étonner que le fait de devoir estimer des paramètres diminue les degrés de libertés. En effet un quantile d'ordre donné diminuant avec les degrés de liberté la valeur critique en est abaissée par rapport à une même situation où θ serait connu. La raison tient au fait que la statistique Q sous-évalue les écarts aux vraies fréquences attendues en leur substituant des fréquences calculées sur l'échantillon lui-même et ceci d'autant plus que le nombre de paramètres à estimer est grand (à la limite, si θ était de dimension $c-1$ on serait simplement en présence d'une reparamétrisation du vecteur $(p_1, p_2, \cdots, p_{c-1})$ et l'on prendrait $\widehat{p}_j = N_j/n$ réduisant ainsi l'expression de Q à 0).

L'application du théorème à l'ajustement dans une famille de lois est immédiate. On procède comme en section 10.4.1 en opérant un découpage en classes, mais ici les probabilités p_k associées aux intervalles $]a_{k-1}, a_k]$ dépendent du paramètre θ. Il s'agit alors d'exprimer chaque p_k comme une fonction de θ et d'en déduire l'estimateur du MV comme indiqué ci-dessus. Remarquons bien que cet estimateur n'est pas, hélas, celui que l'on obtient directement de la façon classique sur la base des observations $X_1, X_2, \cdots, X_n$. Illustrons cela par un exemple.

Exemple 10.2 Soit à tester l'hypothèse que les observations proviennent d'une loi de Poisson $\mathcal{P}(\lambda)$. Les observations au-delà de 3 étant rares, supposons que l'on effectue un découpage en 4 classes : $\{0\}, \{1\}, \{2\}$ et $\{3$ et plus$\}$. Les probabilités associées à ces classes sont, respectivement :

$$e^{-\lambda}\ ,\ e^{-\lambda}\lambda\ ,\ e^{-\lambda}\frac{\lambda^2}{2} \text{ et } 1-(e^{-\lambda}+e^{-\lambda}\lambda\ +\ e^{-\lambda}\frac{\lambda^2}{2})\,.$$

Soit n_i, $i = 1, \cdots, 4$, les fréquences observées dans les 4 classes. L'estimation $\widehat{\lambda}$ du MV approprié est obtenue en maximisant la fonction de vraisemblance de λ suivante :

$$L(\lambda) = \frac{n!}{n_1!n_2!\cdots n_c!}[e^{-\lambda}]^{n_1}[e^{-\lambda}\lambda]^{n_2}[e^{-\lambda}\frac{\lambda^2}{2}]^{n_3}[1-(e^{-\lambda}+e^{-\lambda}\lambda\ +\ e^{-\lambda}\frac{\lambda^2}{2})]^{n_4}\,.$$

Cette fonction de λ n'est pas simple et il faut recourir à un algorithme d'optimisation numérique. Il est clair que la solution est différente de celle de l'estimateur du MV classique fondé sur les observations brutes $x_1, x_2, \cdots, x_n$ et égal à la moyenne $\overline{x}$ des observations. Prenons les 100 observations suivantes :

valeurs	0	1	2	3	4	5
fréquences	38	40	12	7	2	1

Le maximum de $L(\lambda)$ obtenu par un logiciel mathématique est 9,68 alors que la moyenne des observations est 9,8.

Notons que si n_4 est petit ces valeurs sont proches car les fonctions à maximiser sont similaires. En effet pour $L(\lambda)$ il faut maximiser

$$e^{-\lambda(n_1+n_2+n_3)}\lambda^{n_2+2n_3}[1-(e^{-\lambda}+e^{-\lambda}\lambda\ +\ e^{-\lambda}\frac{\lambda^2}{2})]^{n_4}$$

alors que la fonction de vraisemblance classique est proportionnelle à :

$$\prod_{i=1}^{n} e^{-\lambda}\lambda^{\sum_{i=1}^{n} x_i} = e^{-\lambda(n_1+n_2+n_3+n_4'+n_5'\cdots)}\lambda^{n_2+2n_3+3n_4'+4n_5'+\cdots}$$

où n_4', n_5', etc. sont les fréquences observées des valeurs $3, 4$, etc. ■

Cet exemple simple illustre le fait que l'estimation appropriée du maximum de vraisemblance est généralement difficile. Pour tester que la loi mère est gaussienne le problème est encore plus complexe. La probabilité p_k associée à l'intervalle $]a_{k-1}, a_k]$ est égale à $\Phi(\frac{a_k-\mu}{\sigma}) - \Phi(\frac{a_{k-1}-\mu}{\sigma})$, où Φ est la fonction de répartition de la loi de Gauss centrée-réduite, ce qui complique fortement la fonction de vraisemblance $L(\mu, \sigma^2)$. En pratique on utilise l'estimation classique du paramètre qui est d'autant plus proche de l'estimation appropriée que le découpage en classes est fin (mais avec les limitations qui demeurent, à savoir que les fréquences attendues ne descendent pas en dessous de 5). Chernoff et Lehmann (1954) ont montré que, dans ce cas, la statistique Q ne suit pas une loi du khi-deux mais une loi encadrée par les lois $\chi^2(c-r-1)$ et $\chi^2(c-1)$. On se rapproche donc du cas où les fréquences attendues sont parfaitement connues du fait que l'estimateur usuel du MV est plus efficace. En gardant $c-r-1$ degrés de liberté, comme le font les praticiens et la plupart des logiciels, on effectue un test anti-conservateur (i.e. de niveau réel supérieur au niveau nominal) puisque le quantile est inférieur à ce qu'il devrait être. Une procédure assurément conservatrice, mais souvent trop, consisterait à prendre le quantile sur la loi $\chi^2(c-1)$.

Si le nombre de classes c est assez élevé, la différence entre les quantiles sera peu sensible, sachant que presque toutes les familles paramétriques courantes ont un paramètre à une ou deux dimensions ($r \leq 2$).

Test de Kolmogorov-Smirnov

En pratique on adapte le test vu en section 10.4.1 en calculant la statistique :

$$\widetilde{D}_n = \sup_{x\in\mathbb{R}} |F_n(x) - F(x;\widehat{\theta})|$$

où $F(x;\theta)$ est la fonction de répartition pour la famille paramétrique considérée et $\widehat{\theta}$ est l'estimateur usuel du maximum de vraisemblance. Mais alors la loi de $\widetilde{D}_n$ sous H_0 n'est plus indépendante de la vraie loi et il faut étudier chaque cas de famille séparément. Fort heureusement le fait d'utiliser la valeur critique

propre à la statistique D_n de la section 10.4.1 conduit à un test conservateur. Ceci découle de raisons semblables à celles invoquées pour la statistique du khi-deux, à savoir que $\widetilde{D}_n$ tend à sous-estimer l'écart-réel et devrait être rejetée à des valeurs critiques inférieures.

Remarques diverses

1. Il existe des tests spécifiques à chaque famille qui, de ce fait, sont en principe plus puissants que les tests généraux précédents. Citons en particulier le *test de normalité de Shapiro-Wilk* (1965) fondé sur la corrélation entre les statistiques d'ordre et leurs espérances mathématiques sous hypothèse gaussienne.

2. Souvent on n'a pas d'idée préconçue de modèle et l'on recherche, parmi les modèles courants, celui qui est le plus proche des observations. L'avantage d'une procédure générale telle que celles présentées ci-dessus est qu'elle fournit le même critère pour comparer plusieurs modèles, les statistiques D_n ou Q tenant lieu de distance à minimiser. Paradoxalement, on ne souhaite pas avoir un test trop puissant, car on se contente de s'assurer que le modèle le plus proche est accepté par le test d'adéquation.

3. Il existe aussi des méthodes graphiques telle que la droite de Henri pour l'hypothèse gaussienne et sa version générale non-paramétrique du *QQ-plot* qui est le graphe de la variation des quantiles empiriques en fonction des quantiles théoriques sous H_0. Ces méthodes ont l'avantage de mettre en évidence les zones de forte déviation par rapport au modèle supputé et donc d'orienter, le cas échéant, la recherche d'un meilleur modèle. L'inconvénient est que le jugement d'acceptabilité repose sur une appréciation graphique et reste fortement subjectif.

4. Le test du khi-deux s'étend aisément au test d'égalité de deux lois (ou des distributions de deux populations). Il suffit d'utiliser un découpage en classes et de se ramener ainsi à la comparaison de deux lois multinomiales exposée en section 10.2. Il existe également une version à deux échantillons du test de Kolmogorov-Smirnov fondé sur la statistique

$$D^*_{n,m} = \sup_{x\in\mathbb{R}} |F_n(x) - F_m(x)|$$

où F_n et F_m sont les fonctions de répartition empiriques des échantillons de tailles respectives n et m. Quand n et m tendent vers l'infini on a :

$$P((\tfrac{1}{n} + \tfrac{1}{m})^{-1/2} D_{n,m} < x) \simeq 1 - 2e^{-2x^2}$$

qui est la même probabilité que pour $\sqrt{n}D_n < x$ dans le cas d'un seul échantillon. Ceci permet de construire un test approché. Il existe également des tables pour les faibles valeurs de n et m.

10.5 Tests non paramétriques sur des caractéristiques de lois

10.5.1 Introduction

Il y a une certaine ambiguïté en ce qui concerne les tests sur le terme de non paramétrique auquel les anglo-saxons préfèrent parfois celui de «distribution free» pour qualifier les **procédures valides quelle que soit la loi mère**. En particulier la loi F peut alors être totalement non spécifiée ce qui permet de parler de procédures non paramétriques. Une autre ambiguïté vient du fait que l'on assimile généralement les tests non paramétriques aux tests fondés sur les rangs des observations. Il est vrai que les tests de rangs sont, par essence, applicables indépendamment de la nature de la loi mère et qu'ils offrent de nombreuses possibilités, mais il existe d'autres tests de type «distribution free» comme, par exemple, les tests d'ajustement du khi-deux et de Kolmogorov-Smirnov vus en section 10.4.1 ou certains des tests qui vont suivre.

Nous ne présenterons que les tests les plus courants pour illustrer la philosophie générale de l'approche non paramétrique. Étant donné, malgré tout, la place importante des rangs dans cette approche, nous donnons tout d'abord quelques propriétés les concernant.

10.5.2 Les statistiques de rang

On considère un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ de loi F. Pour des réalisations $(x_1, x_2, \ldots, x_n)$, le *rang* r_i d'une valeur x_i est la position qu'elle occupe quand les valeurs sont rangées dans l'ordre croissant. A tout vecteur de réalisations on peut donc associer le vecteur des rangs $(r_1, r_2, \ldots, r_n)$ qui consiste en une permutation des nombres $\{1, 2, \ldots, n\}$. Par exemple, avec $n = 5$, au vecteur (8,2 ; 7,4 ; 9,2 ; 5,1 ; 6,7) on associe le vecteur des rangs $(4, 3, 5, 1, 2)$. Cette fonction appliquée à $(X_1, X_2, \ldots, X_n)$ procure les *statistiques de rang* $(R_1, R_2, \ldots, R_n)$. La v.a. R_i sera appelée le rang de X_i. Notons que si X_i est la statistique d'ordre k alors R_i est égal à k. On supposera que F est continue afin d'ignorer, pour l'heure, le problème des valeurs identiques.

La proposition suivante indique que les statistiques de rang ont une loi conjointe indépendante de F et, par conséquent, que toute inférence fondée sur les rangs sera de nature non paramétrique. Qui plus est, cette loi est parfaitement déterminée et permettra d'établir les distributions d'échantillonnage des statistiques de test reposant sur les rangs.

Proposition 10.1 *La loi de* $(R_1, R_2, \ldots, R_n)$ *est la loi uniforme sur l'ensemble des* $n!$ *permutations des nombres* $\{1, 2, \ldots, n\}$.

Ce résultat découle du fait que toutes les v.a. $X_1, X_2, \ldots, X_n$ sont *échangeables*, c'est-à-dire que toute permutation des composantes de $(X_1, X_2, \ldots, X_n)$

a la même loi conjointe que $(X_1, X_2, \ldots, X_n)$. On en déduit également les lois marginales des rangs.

Proposition 10.2 *Pour tout $i = 1, \ldots, n$ le rang R_i suit une loi discrète uniforme sur $\{1, 2, \ldots, n\}$.*

Ainsi (voir section 4.1.1) :

$$E(R_i) = \frac{n+1}{2}$$
$$V(R_i) = \frac{n^2-1}{12}\,.$$

De plus on démontre que, pour tout i et tout j distincts,

$$cov(R_i, R_j) = -\frac{n+1}{12}.$$

Outre leur non dépendance vis-à-vis de la loi mère les statistiques de rang ont l'avantage de pouvoir s'appliquer lorsque les données sont peu précises et même simplement ordinales. Ceci est souvent le cas dans les tests psychologiques et dans les études de comportement d'achat ou de consommation, où les sujets sont amenés à exprimer des préférences ou à effectuer des classements. De plus ces statistiques sont peu sensibles aux valeurs extrêmes ou aberrantes.

10.5.3 Tests sur moyenne, médiane et quantiles

Test sur la moyenne

En section 8.2.1 on a pu voir que, si μ est la moyenne de la loi, on a :

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1)$$

si n est assez grand, pourvu que la loi mère admette une variance. En conséquence le test de Student de la section 9.7.1 fournit un test approché pour une hypothèse du type $H_0 : \mu = \mu_0$.

Si l'échantillon est trop petit pour garantir une bonne approximation, ou si la loi mère peut produire des valeurs extrêmes (queues de distribution allongées), ou s'il y a risque de présence de valeurs aberrantes, il sera préférable de recourir à un test non paramétrique concernant la médiane $\widetilde{\mu}$ de la loi.

Test sur la médiane : le test du signe

Ce test est le dual de la procédure d'intervalle de confiance pour la médiane vue en section 8.3. Comme alors, nous supposons simplement que la fonction de répartition de la loi mère F est continue et strictement croissante pour garantir l'unicité de la médiane. L'hypothèse nulle est $H_0 : \widetilde{\mu} = \widetilde{\mu}_0$ avec une alternative

soit unilatérale soit bilatérale. Pour l'échantillon aléatoire $X_1, X_2, \ldots, X_n$ la statistique de test est $\widetilde{N}$, le nombre de ces v.a. inférieures ou égales à $\widetilde{\mu}_0$. Sous H_0, pour tout i on a $P(X_i \leq \widetilde{\mu}_0) = \frac{1}{2}$, donc $\widetilde{N}$ suit une loi $\mathcal{B}(n, \frac{1}{2})$.

Pour le cas bilatéral on rejette H_0 si $\widetilde{N}$ prend une valeur soit trop grande soit trop petite, les valeurs critiques devant être choisies de façon conservatrice en raison du caractère discret de la loi binomiale (plus commodément on pourra se contenter d'indiquer la P-valeur de la valeur observée de la statistique). Considérons le test unilatéral $H_0 : \widetilde{\mu} = \widetilde{\mu}_0$ (ou $H_0 : \widetilde{\mu} \leq \widetilde{\mu}_0$) versus $H_1 : \widetilde{\mu} \geq \widetilde{\mu}_0$. Sous H_1, $\widetilde{\mu}_0$ est inférieure à la médiane et la probabilité d'observer une valeur inférieure à $\widetilde{\mu}_0$ est inférieure à $\frac{1}{2}$. $\widetilde{N}$ suit donc une loi $\mathcal{B}(n, p)$ où $p < \frac{1}{2}$ et l'on rejettera l'hypothèse nulle lorsque $\widetilde{N}$ prendra une valeur trop petite. A l'inverse pour $H_1 : \widetilde{\mu} \leq \widetilde{\mu}_0$ on rejettera H_0 lorsque $\widetilde{N}$ prendra une valeur trop grande.

Ce test est appelé *test du signe* car, en retranchant préalablement $\widetilde{\mu}_0$ à chaque observation, $\widetilde{N}$ devient le nombre de valeurs négatives (ou nulles).

Étant identique à un test sur le paramètre p d'une loi de Bernoulli le test est sans biais. Le calcul de sa puissance repose sur $p = P(X_i \leq \widetilde{\mu}_0)$ qui dépend de F choisi dans l'alternative H_1. Étant donné que l'information initiale est ramenée à une information binaire on ne peut s'attendre à un test très puissant. Ceci est la contrepartie de sa validité très générale.

F étant supposée continue la probabilité qu'une valeur soit exactement égale à $\widetilde{\mu}_0$ est nulle. Toutefois, en raison du caractère discret de toute mesure pratique, il se peut qu'une ou plusieurs valeurs soient égales à $\widetilde{\mu}_0$ et donc inclassables dans la procédure. On dit avoir affaire à un *problème d'ex aequo.* Le remède recommandé par Lehmann (1975) consiste à ignorer ces valeurs, diminuant d'autant la taille de l'échantillon.

Le test du signe s'applique également au cas des **échantillons appariés**. Pour fixer les idées prenons le cas de n individus observés avant et après un traitement. Soit (X_i, Y_i), $i = 1, \cdots, n$, les n couples d'observations correspondantes et $p = P(Y_i > X_i)$. L'hypothèse nulle que la distribution est identique avant et après un traitement, à savoir qu'il n'y a pas d'effet du traitement, implique que $p = \frac{1}{2}$. La statistique est alors le nombre de v.a. $Y_i - X_i$ négatives, les valeurs nulles devant être écartées. Il est intéressant de noter que la procédure s'applique au cas discret (le test étant conditionnel aux différences non nulles) et lorsque l'information sur chaque couple est un simple classement.

Le test s'étend aisément à un test portant sur un quantile en remplaçant la valeur $p = \frac{1}{2}$ dans H_0 par l'ordre du quantile considéré.

10.5.4 Tests de localisation de deux lois

On considère ici qu'on est en présence de deux lois dont les fonctions de répartition F_1 et F_2 ont la même forme mais peuvent être localisées différemment. En d'autres termes leurs graphes sont identiques à une translation près.

Nous supposons que ces lois sont continues, alors leurs densités sont également translatées l'une par rapport à l'autre. Une telle situation n'est pas rare, notamment pour les échantillons appariés qui feront l'objet d'une attention particulière. Remarquons d'ailleurs que la condition d'égalité des variances imposée en section 9.7.3 pour tester l'égalité des moyennes de deux lois de Gauss induit une situation de ce type.

Mathématiquement le *modèle de localisation* (ou modèle de position) s'écrit :

$$F_2(x) = F_1(x - \delta), \text{ pour tout } x \in \mathbb{R},$$

où δ est une constante inconnue caractérisant le décalage des deux lois. Si δ est positif le graphe de la densité de la deuxième loi est translaté à droite de celui de la première, il est à gauche si δ est négatif. Le test porte sur l'hypothèse nulle d'identité de ces lois ce qui équivaut à :

$$H_0 : \delta = 0 \quad \text{vs.} \quad H_1 : \delta \neq 0.$$

On peut également envisager des tests unilatéraux.

Nous présentons en premier lieu le test de Wilcoxon qui illustre bien l'usage des rangs dans l'approche non paramétrique.

Test de Wilcoxon ou de Mann-Whitney

Ce test a été proposé initialement par Wilcoxon (1945). Par la suite Mann et Whitney (1947) ont proposé une forme équivalente qui permit de préciser ses propriétés.

Soit deux **échantillons indépendants** $X_1, X_2, \ldots, X_{n_1}$ et $Y_1, Y_2, \ldots, Y_{n_2}$ issus respectivement de chaque loi. Considérons la fusion des $n_1 + n_2$ valeurs en un seul échantillon et les rangs associés à celui-ci. La statistique de test de Wilcoxon est la somme des rangs de l'un des échantillons initiaux. Il est plus rapide de choisir celui de plus petite taille et nous supposerons qu'il s'agit du premier (soit $n_1 \leq n_2$), notant alors la somme de ses rangs T_{n_1}. La valeur minimale pour T_{n_1} est atteinte lorsque toutes les réalisations $x_1, x_2, \ldots, x_{n_1}$ sont situées à gauche des réalisations $y_1, y_2, \ldots, y_{n_2}$ sur la droite réelle et elle vaut $1 + 2 + \cdots + n_1 = \frac{1}{2}n_1(n_1 + 1)$. La valeur maximale est atteinte lorsque toutes les observations x_i sont situées à droite des observations y_j sur la droite réelle et elle vaut :

$$(n_2 + 1) + (n_2 + 2) + \cdots + (n_2 + n_1) = n_1 n_2 + \frac{1}{2}n_1(n_1 + 1).$$

Intuitivement on est enclin à rejeter H_0 lorsque la valeur de T_{n_1} s'approche de l'un ou l'autre de ces extrêmes (mais un seul d'entre eux pour un cas unilatéral). Pour déterminer les valeurs critiques il est nécessaire d'établir la loi de cette statistique sous H_0. Cela peut être fait par des considérations combinatoires, lesquelles ont conduit à la construction de tables bien répandues. Nous montrons sur un exemple la démarche utilisée.

Exemple 10.3 Soit les résultats suivants :

	échant. 1 $n_1 = 4$				échant. 2 $n_2 = 5$				
valeurs	15	24	12	10	35	25	20	29	16
rangs	*3*	*6*	*2*	*1*	*9*	*7*	*5*	*8*	*4*

La statistique de test T_{n_1} prend la valeur 12. Calculons la P-valeur correspondante pour un test bilatéral en établissant la loi de T_{n_1} sous H_0, en partant des valeurs extrêmes 10 et 30.

Sous H_0 toutes les 9 v.a. sont de même loi et la proposition 10.1 s'applique : les 9 ! permutations des rangs sont équiprobables. Pour calculer $P(T_{n_1} = 10)$ il faut dénombrer les permutations de rangs aboutissant à une somme 10 pour les rangs du premier échantillon. Pour cela il faut que ces rangs soient {1,2,3,4} dans un ordre quelconque soit 4 ! possibilités. Pour chacune de ces possibilités on peut permuter les 5 rangs restant pour le deuxième échantillon. Il y a donc en tout 4! 5! cas possibles, d'où :

$$P(T_{n_1} = 10) = \frac{4!\,5!}{9!} = 0{,}0079365\,.$$

Pour l'événement $(T_{n_1} = 30)$ il faut que les rangs du premier échantillon soient une permutation sur {6,7,8,9} ce qui conduit à la même probabilité. Examinons maintenant l'événement $(T_{n_1} = 11)$. Il n'y a toujours qu'une possibilité pour la liste des rangs du premier échantillon, soit {1,2,3,5}, donc encore la même probabilité que ci-dessus. Ceci vaut également pour $(T_{n_1} = 29)$. Enfin, pour obtenir $(T_{n_1} = 12)$, deux listes sont possibles : {1,2,4,5} et {1,2,3,6}. De même pour $(T_{n_1} = 28)$ on a deux listes : {5,6,8,9} et {4,7,8,9}. Donc $P(T_{n_1} = 12) = P(T_{n_1} = 28) = 2\frac{4!\,5!}{9!} = 0{,}015873$. Finalement :

$$P(T_{n_1} \leq 12) = P(T_{n_1} \geq 28) = 4\,\frac{4!\,5!}{9!} \simeq 0{,}032$$

et la P-valeur, pour une alternative bilatérale, est 0,064. Si l'on se fixe comme niveau $\alpha = 0{,}05$ on doit accepter H_0. Pour une alternative unilatérale, par exemple que le graphe de la densité de la loi mère du deuxième échantillon soit translaté à droite de celui du premier, i.e. $H_1 : \delta > 0$, alors la P-valeur serait égale à 0,032 et il faudrait rejeter H_0, donc considérer qu'il y a bien translation à droite. ■

Cet exemple est instructif sur plusieurs aspects. Tout d'abord on voit que le point crucial pour déterminer la loi de T_{n_1} est de dénombrer les façons d'obtenir un total donné en choisissant k entiers parmi les n premiers entiers (ici $k = n_1$ et $n = n_1 + n_2$). Ceci est un problème de combinatoire résolu par une relation de récurrence en partant du plus petit total. De plus, ce dénombrement est identique en partant du plus grand total et, par conséquent, la loi de la statistique est symétrique. Ainsi on peut aisément construire des tables de

valeurs critiques, lesquelles doivent être conservatrices vu le caractère discret de la loi. Enfin on voit l'intérêt de la procédure dans le cas de petits échantillons.

Pour les grandes tailles d'échantillons (en fait $n_1 > 10$ et $n_2 > 10$ suffisent) on peut utiliser une approximation gaussienne découlant du comportement asymptotique de T_{n_1} sous H_0. Pour cela il faut utiliser la moyenne et la variance de cette statistique sous H_0 :

$$E(T_{n_1}) = \frac{n_1(n_1+n_2+1)}{2}$$
$$V(T_{n_1}) = \frac{n_1 n_2(n_1+n_2+1)}{12}\,.$$

L'espérance mathématique, en raison de la symétrie de la loi, est simplement la demi-somme des deux valeurs extrêmes données plus haut. Pour établir la variance on peut utiliser les formules générales sur les moments des rangs indiquées à la suite de la proposition 10.2, de la façon suivante.

Comme $T_{n_1} = \sum_{i=1}^{n_1} R_i$, on a, par extension de la formule sur la variance d'une somme de deux v.a. non indépendantes vue en section 3.5,

$$V(T_{n_1}) = \sum_{i=1}^{n_1} V(R_i) + 2\sum_{i<j} cov(R_i, R_j)$$

où la somme $\sum_{i<j}$ est à effectuer sur tous les $\frac{n_1(n_1-1)}{2}$ couples (R_i, R_j) du premier échantillon tels que $i < j$. Comme $V(R_i) = \frac{(n_1+n_2)^2-1}{12}$ et $cov(R_i, R_j) = -\frac{n_1+n_2+1}{12}$ quels que soient i et j, on obtient finalement :

$$\begin{aligned} V(T_{n_1}) &= n_1\frac{(n_1+n_2)^2-1}{12} - 2\,\frac{n_1(n_1-1)}{2}\,\frac{(n_1+n_2+1)}{12} \\ &= \frac{n_1(n_1+n_2+1)}{12}\,[(n_1+n_2-1)-(n_1-1)] \\ &= \frac{n_1 n_2(n_1+n_2+1)}{12}. \end{aligned}$$

La statistique U proposée par Mann et Whitney est la suivante : pour chaque Y_j on compte le nombre d'observations X_i qui lui sont supérieures puis on totalise ces nombres pour $j = 1, \cdots, n_2$. En posant :

$$Z_{ij} = \begin{cases} 1 & \text{si } X_i > Y_j \\ 0 & \text{si } X_i < Y_j \end{cases},$$

cette statistique s'écrit :

$$U = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} Z_{ij}\,.$$

On montre (voir exercices) que U est égale à T_{n_1} à une constante près :

$$U = T_{n_1} - \frac{n_1(n_1+1)}{2}\,.$$

Ainsi on peut aussi établir la procédure et les résultats précédents à partir de U. Certaines tables sont d'ailleurs données en fonction de cette statistique.

Nous avons ignoré jusqu'ici le problème des *ex aequo,* c'est-à-dire lorsque deux (ou plusieurs) valeurs sont égales et ne peuvent être rangées. Le remède le plus simple est celui des *rangs moyens* qui consiste à attribuer à chacune de ces valeurs la moyenne des rangs qu'elles auraient totalisés si elles avaient été différentiées (si, par exemple, il y a deux valeurs identiques après la sixième valeur, chacune reçoit le rang 7,5 ; trois valeurs identiques recevraient le rang 8). En théorie ceci nécessite un correctif pour la statistique de test mais qui reste mineur si les ex aequo ne sont pas trop nombreux. Une autre méthode plus efficace mais plus lourde consiste à attribuer les rangs de façon aléatoire.

De nombreuses études, soit asymptotiques, soit par simulations pour des tailles d'échantillons réduites, ont été effectuées pour étudier la puissance du test en fonction de divers types de lois mères. Ceci est notamment vrai pour le cas de lois de Gauss qui mène à une comparaison avec le test de Student classique. Asymptotiquement le rapport de la puissance du test de Wilcoxon à celle du test de Student est de 0,96, cette valeur étant pratiquement atteinte avec des tailles d'échantillons de l'ordre de 50. Hodges et Lehmann (1956) ont établi que le rapport asymptotique ne descend pas au-dessous de 0,86 quelle que soit la loi.

Ces résultats justifient certainement l'usage répandu de ce test. Cependant la condition d'un modèle de localisation est une restriction importante. Cette condition peut toutefois être assouplie. Si l'hypothèse alternative est de la forme $F_1(x) > F_2(x)$ pour tout x ou $F_1(x) < F_2(x)$, c'est-à-dire que les graphes restent totalement décalés, les propriétés du test sont globalement conservées.

Echantillons appariés : test des rangs signés

Ce test, également dû à Wilcoxon, est un dérivé du précédent. Il repose sur le fait que, si les lois mères sont identiques, la loi des différences $X_i - Y_i$ doit être symétrique par rapport à 0. Ainsi on s'attend à ce que les rangs des différences absolues $|X_i - Y_i|$ se partagent équitablement pour les différences positives et pour les différences négatives. Ayant rangé les $|X_i - Y_i|$ par valeurs croissantes, la statistique de test est la somme T^+ des rangs associés aux différences $X_i - Y_i$ positives.

Soit les v.a. Z_i , $i = 1, \cdots, n$, définies par

$$Z_i = \begin{cases} 1 & \text{si } X_i - Y_i > 0 \\ 0 & \text{si } X_i - Y_i < 0 \end{cases}\,.$$

Alors T^+ est égal à $\sum_{i=1}^{n} iZ_i$. Sous H_0, Z_i suit une loi de Bernoulli $\mathcal{B}(\frac{1}{2})$ ce qui permet d'établir la loi de T^+. Ses valeurs possibles sont $0, 1, \cdots, \frac{n(n+1)}{2}$, la valeur 0 étant atteinte lorsqu'il n'y a aucune différence positive et la valeur $\frac{n(n+1)}{2}$ lorsque toutes les différences sont positives. Dans le premier cas cela porte à croire que la (densité de la) loi mère des X_i est décalée à gauche de celle des Y_i et, dans le deuxième cas, qu'elle est décalée à droite. Ceci oriente donc le sens du rejet pour un test unilatéral.

On trouve aisément des tables de valeurs critiques et l'on peut utiliser une approximation gaussienne dès que $n > 30$ en utilisant la moyenne et la variance de T^+. Comme $E(Z_i) = \frac{1}{2}$, $V(Z_i) = \frac{1}{4}$ et que les Z_i sont mutuellement indépendantes (comme fonctions respectives des paires (X_i, Y_j)), on a :

$$E(T^+) = \frac{1}{2}\sum_{i=1}^{n} i = \frac{n(n+1)}{4}$$

$$V(T^+) = \frac{1}{4}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}$$

sachant que $1^2 + 2^2 + \cdots + n^2 = \frac{1}{6}n(n+1)(2n+1)$.

Sous H_1 les Z_i ont une loi $\mathcal{B}(p)$ avec $p = P(X_i - Y_i > 0)$ et la puissance peut être calculée en fonction de p, lequel dépend toutefois de la loi mère des $X_i - Y_i$. Dans un problème de localisation le test est sans biais et est plus puissant que le test du signe mentionné en section 10.5.3.

Le test des rangs signés est parfois employé comme alternative au test du signe dans le cas d'un seul échantillon, mais cette pratique est contestable en raison de la **condition forte de symétrie de la loi mère**, peu réaliste dans une approche non paramétrique.

Test de la médiane

Ce test opère dans le même cadre que le test de Wilcoxon pour deux échantillons indépendants et relève du même esprit que le test du signe. On détermine la médiane des $n_1 + n_2$ valeurs fusionnées et l'on considère le nombre d'observations du premier échantillon inférieures à cette médiane globale. Désignons par $\widetilde{N}_1$ la statistique correspondante. Supposons par commodité que $n_1 + n_2$ est pair et posons $n_1 + n_2 = 2r$, de façon qu'il y ait exactement r observations à gauche comme à droite de la médiane globale (celle-ci peut être indifféremment n'importe quelle valeur entre les deux observations les plus centrales). Sous l'hypothèse nulle d'identité des deux lois on s'attend à une valeur de $\widetilde{N}_1$ proche de $n_1/2$ et l'on rejettera donc H_0 si la réalisation de cette statistique s'éloigne trop de $n_1/2$. Étant donné que l'échantillon fusionné a été divisé en deux parties de taille égale et que, sous H_0, toutes les v.a. sont i.i.d., la loi de $\widetilde{N}_1$, nombre d'observations parmi n_1 observations appartenant à l'ensemble

des r plus petites valeurs, correspond à la définition même (voir section 4.1.5) d'une loi hypergéométrique $\mathcal{H}(2r, r, n_1)$. On en déduit immédiatement :

$$E(\widetilde{N}_1) = \frac{n_1}{2} \qquad \text{et} \qquad V(\widetilde{N}_1) = \frac{n_1 n_2}{4(n_1 + n_2 - 1)}\,,$$

valeurs à utiliser pour une approximation gaussienne dès lors que n_1 et n_2 sont assez grands. Si $n_1 + n_2$ est impair on montre, en posant $n_1 + n_2 = 2r + 1$, que $\widetilde{N}_1$ suit une loi $\mathcal{H}(2r + 1, r, n_1)$. Ainsi, dans ce cas :

$$E(\widetilde{N}_1) = \frac{n_1(n_1 + n_2 - 1)}{2(n_1 + n_2)} \qquad \text{et} \qquad V(\widetilde{N}_1) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{4(n_1 + n_2)^2}\,.$$

Comme on peut s'y attendre ce test est moins puissant que le test de Wilcoxon du fait qu'il ignore les rangs des observations (ainsi, pour la comparaison dans le cas de lois mères gaussiennes, le rapport asymptotique de sa puissance à celle du test de Student tombe à 0,63 contre 0,96 pour le test de Wilcoxon). En contrepartie sa portée dépasse le seul modèle de localisation. Il peut être appliqué, par exemple, comme test d'égalité des médianes des deux lois. Par ailleurs il offre un substitut au test de Wilcoxon si le nombre d'ex aequo est important suite à une forte discrétisation des données recueillies. Son pendant pour deux échantillons appariés est le test du signe décrit plus haut.

Estimateur de Hodges-Lehmann du décalage des deux lois

On cherche à estimer le paramètre δ qui caractérise le modèle de localisation, à savoir tel que :

$$F_2(x) = F_1(x - \delta), \text{ pour tout } x \in \mathbb{R},$$

où F_1 est la fonction de répartition des X_i et F_2 celle des Y_j. On peut estimer ponctuellement δ par la différence $\overline{y} - \overline{x}$ des moyennes observées dans chaque échantillon et fournir un intervalle de confiance approché avec la procédure classique de Student (supposant que ces lois admettent une moyenne et une variance). Toutefois, pour de petits échantillons et/ou en présence de valeurs extrêmes, il est souhaitable de disposer de procédures plus fiables. Hodges et Lehmann (1963) ont proposé une approche en relation avec les tests non paramétriques que nous exposons dans le cas du test de Wilcoxon-Mann-Whitney, le principe étant identique pour d'autres tests. Nous nous intéressons particulièrement à l'intervalle de confiance.

Pour une valeur arbitraire δ_0, on détermine la valeur $t_{n_1}(\delta_0)$ prise par la statistique de Wilcoxon pour la série fusionnée :

$$x_1, \ldots, x_i, \ldots, x_{n_1},\, y_1 - \delta_0, \ldots, y_j - \delta_0, \ldots, y_{n_2} - \delta_0\,.$$

En vertu de l'équivalence test-IC (voir section 9.8) on prend comme IC de niveau $1-\alpha$ pour δ l'ensemble des valeurs δ_0 telles que $t_{n_1}(\delta_0)$ reste à l'intérieur

des valeurs critiques au niveau α du test. Ceci est justifié par le fait que, si δ_0 est la vraie valeur, les X_i et les $Y_j - \delta_0$ ont la même loi et la statistique $T_{n_1}(\delta_0)$ suit alors la loi parfaitement déterminée sous H_0 pour des échantillons de taille n_1 et n_2.

Faire varier δ_0 peut être fastidieux mais cela n'est pas nécessaire car on montre que cet IC peut être obtenu de façon simple et directe comme suit. On a vu plus haut que la statistique T_{n_1} de Wilcoxon a, sous H_0, une loi symétrique sur l'ensemble des entiers de $\frac{1}{2}n_1(n_1+1)$ à $n_1n_2 + \frac{1}{2}n_1(n_1+1)$. Soit k_α le plus petit entier tel que, sur cette loi, la probabilité associée à l'intervalle :

$$[\frac{1}{2}n_1(n_1+1) + k_\alpha\,,\; n_1n_2 + \frac{1}{2}n_1(n_1+1) - k_\alpha]$$

soit au moins égale à $1-\alpha$. On considère alors la série des n_1n_2 valeurs $y_j - x_i$ pour tous i et j. L'intervalle de confiance de niveau (conservateur) $1-\alpha$ s'obtient en prenant comme bornes les statistiques d'ordres k_α et $n_1n_2 - k_\alpha + 1$ de cette série de valeurs.

Pour l'estimation ponctuelle on choisit la valeur δ_0 telle que $t_{n_1}(\delta_0)$ coïncide avec la valeur de probabilité maximale sur la loi de référence. On montre que cette estimation est simplement la médiane de la série des $y_j - x_i$. Cette méthode fournit des estimateurs sans biais.

Exemple 10.4 Reprenons l'exemple 10.3. On a vu que, pour $n_1 = 4$ et $n_1 = 5$, la statistique T_{n_1} peut prendre les valeurs entières de 10 à 30 et que

$$P(T_{n_1} \leq 11) + P(T_{n_1} \geq 29) = 0{,}032\,.$$

Donc $P(12 \leq T_{n_1} \leq 28) = 1-0{,}032$ alors que $P(13 \leq T_{n_1} \leq 27) = 1-0{,}064$. Pour $\alpha = 0{,}05$ on a donc $k_{0,05} = 2$. La série des $y_j - x_i$ ordonnée est :

$$-8, -4, 1, 1, 4, 5, 5, 6, 6, 8, 10, 10, 13, 14, 15, 15, 17, 18, 19, 20$$

donc l'intervalle de confiance est $[-4\,;\,19]$ et l'estimateur ponctuel est 9 (en prenant pour médiane la valeur milieu entre 8 et 10).

L'approche classique donne pour estimation ponctuelle $\overline{y} - \overline{x} = 9{,}75$. L'intervalle de confiance ci-dessus étant en fait de niveau 0,968 , celui de même niveau obtenu par la formule de Student (voir section 7.4.3) avec $t^{(7)}_{0,984} = 2{,}67$ est $[-2{,}7\,;22{,}2]$. Ce dernier a une amplitude plus grande mais les deux intervalles restent assez semblables du fait qu'il n'y a pas de valeurs extrêmes. ■

10.5.5 Test pour la corrélation de Spearman

Nous présentons brièvement ce test pour illustrer encore l'intérêt des procédures reposant sur les rangs. On considère un couple (X, Y) de fonction de répartition conjointe $F_{X,Y}(x, y)$ inconnue et un échantillon de taille n :

$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, issu de cette même loi. On souhaite tester l'hypothèse nulle d'indépendance de ces deux composantes du couple, soit :

$$H_0 : F_{X,Y}(x, y) = F_X(x)F_Y(y) \text{ pour tout } (x, y) \in \mathbb{R}^2 ,$$

où F_X et F_Y sont les lois marginales des composantes. L'alternative H_1 est la négation de H_0, à savoir qu'il existe au moins un couple de valeurs (x, y) tel que $F_{X,Y}(x, y) \neq F_X(x)F_Y(y)$.

En 1904 Spearman a proposé comme statistique de test la *corrélation des rangs* que nous notons R_S. Elle est obtenue simplement en remplaçant, séparément, dans chaque composante les observations par leurs rangs et en calculant sur ces derniers la corrélation linéaire empirique vue en section 9.7.7. Notons que la valeur prise par R_S sera égale à 1 (respectivement -1) si Y est une fonction croissante (respectivement décroissante) de X. La corrélation de rangs a pour intérêt de mettre en évidence des **liens non linéaires**.

Sous H_0 on s'attend à ce que R_s reste proche de zéro. Comme cette statistique repose sur les rangs sa loi ne dépend pas de F_X ni de F_Y. En établissant (voir exercices) que R_s peut aussi s'écrire :

$$R_s = \frac{\sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} ,$$

où R_i est le rang de X_i et S_i celui de Y_i, et en utilisant l'espérance et la variance d'un rang données à la suite de la proposition 10.2, on montre aisément que :

$$E(R_s) = 0$$
$$V(R_s) = \frac{1}{n-1} .$$

Par ailleurs on montre que :

$$\frac{\sqrt{n-2}R_S}{\sqrt{1-R_S^2}}$$

suit approximativement une loi de Student à $n-2$ degrés de liberté, ce qui est à rapprocher du résultat de la section 9.7.7 concernant la corrélation linéaire. Pour les faibles valeurs de n on dispose de tables des valeurs critiques à différents niveaux.

Nous avons vu quelques tests non paramétriques parmi les plus courants. Les tests fondés sur les rangs sont d'une grande variété et font l'objet d'ouvrages spécifiques. Citons par exemple le livre collectif édité par Droesbeke et Fine (1996), celui de Lecoutre et Tassi (1987) ainsi qu'en anglais Lehmann (1975) et Gibbons (1985).

10.6 Exercices

Exercice 10.1 Montrer que le test du khi-deux de la section 10.2 est le test d'égalité des paramètres de deux lois de Bernoulli vu en section 9.7.6 .

Aide : on s'inspirera de la démarche de la section 10.1.4.

Exercice 10.2 Soit un couple $(\mathcal{X}, \mathcal{Y})$ de variables catégorielles, $\mathcal{X}$ avec I catégories notées $\{1, \cdots, i, \cdots, I\}$ et $\mathcal{Y}$ avec J catégories notées $\{1, \cdots, j, \cdots, J\}$. Montrer qu'elles sont indépendantes si et seulement si il y a indépendance entre tous les couples élémentaires (i, j) de catégories croisées.

Aide : Soit $A = \{i_1\}$ un événement sur $\mathcal{X}$ et $B = \{j_1, j_2\}$ un événement sur $\mathcal{Y}$. (i.e. le résultat de l'expérience est la catégorie j_1 ou j_2), montrer que A et B sont indépendants si $\{i_1\}$ est indépendant de $\{j_1\}$ et de $\{j_2\}$ respectivement, puis généraliser.

Exercice 10.3 Dans le contexte du test exact de Fisher (section 10.3.2) montrer que sous l'hypothèse d'indépendance on a :

$$P(N_{11} = n_{11} | n_{1.}, n_{2.}, n_{.1}, n_{.2}) = \frac{n_{1.}!\, n_{2.}!\, n_{.1}!\, n_{.2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}$$

avec les contraintes nécessaires sur n_{11}, n_{12}, n_{21} et n_{22}.
En déduire que la loi de la v.a. N_{11} conditionnellement aux marges est une loi $\mathcal{H}(n, n_{1.}, n_{.1})$.

Aide : on suivra la même démarche que dans la démonstration du même type de la section 9.7.6 .

Exercice 10.4 (Test de McNemar) Soit un échantillon apparié de n couples d'individus. Sur chacun des $2n$ individus on observe une même variable binaire succès/échec. Pour chaque couple on a donc une variable catégorielle à 4 catégories. Les probabilités et les fréquences (entre parenthèses) sont notées selon le tableau ci-après :

Indiv.1 \ Indiv.2	succès	échec	
succès	p_{11} (n_{11})	p_{12} (n_{12})	$p_{1.}$ $(n_{1.})$
échec	p_{21} (n_{21})	p_{22} (n_{22})	$p_{2.}$ $(n_{2.})$
	$p_{.1}$ $(n_{.1})$	$p_{.2}$ $(n_{.2})$	1 (n)

On considère l'hypothèse nulle que la probabilité de succès est la même pour les deux individus d'un couple, i.e. $H_0 : p_{1.} = p_{.1}$.

1. Donner la fonction de vraisemblance pour cette loi multinomiale à 4 catégories sous H_0. (aide : noter que H_0 équivaut à $p_{12} = p_{21}$ et intégrer les contraintes dans la fonction de vraisemblance comme en section 10.1.1)

2. Donner les estimations du MV des p_{ij} sous H_0.

3. En déduire les estimations des fréquences attendues sous H_0 et montrer que la réalisation q de la statistique Q du test du khi-deux pour H_0 est :

$$q = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} .$$

4. Quels sont les degrés de liberté de la loi asymptotique de cette statistique sous H_0 ?

Exercice 10.5 Démontrer la relation entre la statistique U de Mann-Whitney et T_{n_1} de Wilcoxon.

Aide : on considérera les statistiques d'ordre $X_{(1)}, X_{(2)}, \ldots, X_{(n_1)}$ et les rangs correspondants $R_{(1)}, R_{(2)}, \ldots, R_{(n_1)}$. On exprimera alors chacun de ces rangs en fonction du nombre d'observations Y_j inférieures à la statistique d'ordre à laquelle il correspond.

Exercice 10.6 Montrer que la corrélation de Spearman peut s'écrire

$$R_s = \frac{\sum_{i=1}^{n} R_i S_i - \frac{n(n+1)^2}{4}}{\frac{n(n^2-1)}{12}} .$$

Aide : utiliser les formules générales de décentrage :

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} \;\text{ et }\; \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 .$$

Exercices appliqués

Exercice 10.7 Une enquête sur la gêne causée par la proximité d'un aéroport a donné, par sexe, les résultats suivants :

Gêne \ Sexe	Femmes	Hommes	Tous
Aucune	75	35	110
Faible	25	27	52
Moyenne	17	8	25
Forte	3	12	15
	120	82	202

Identifier la situation d'échantillonnage et poser l'hypothèse nulle correspondant à la question informelle : la gêne est-elle identique pour les deux sexes ? Tester cette hypothèse nulle.

Exercice 10.8 Lors d'une enquête auprès de 825 familles ayant eu 3 enfants on a relevé le nombre de garçons dans chaque famille comme suit :

Nombre de garçons	0	1	2	3	Tous
fréquences	71	297	336	121	825

On fait l'hypothèse que les sexes des enfants lors des naissances successives au sein d'une famille sont des variables catégorielles indépendantes et que la probabilité p d'avoir un garçon reste constante. Déterminer en fonction de p la loi du nombre de garçons pour une famille de 3 enfants. Estimer p et tester l'hypothèse de départ.

Aide : on utilisera le test du khi-deux avec l'estimation de p usuelle par le maximum de vraisemblance.

Exercice 10.9 On donne, pour une agglomération, la répartition du nombre de jours sans accident, avec un accident etc., parmi 50 jours d'observation tirés au hasard dans une année :

nombre d'accidents	nombre de jours
0	21
1	18
2	7
3	3
4	1
total	50

Tester que la répartition du nombre quotidien d'accidents suit une loi de Poisson.

Aide : on effectuera le test du khi-deux en regroupant les catégories de façon à ne pas avoir de fréquences inférieures à 5. Pour simplifier on estimera λ par l'estimation usuelle du maximum de vraisemblance.

Exercice 10.10 Dans une enquête auprès de 93 étudiant(e)s sélectionnés au hasard dans une université on pose une question sur le mode de logement avec 4 modalités de réponse : seul (S), dans la famille (F), en couple (C) et autres modes (A). Les résultats obtenus par sexe sont les suivants :

	S	F	C	A
Féminin	12	11	14	12
Masculin	15	6	9	14

Tester l'hypothèse d'indépendance du mode de logement et du sexe.

Exercice 10.11 Un échantillon de 490 utilisateurs de téléphones portables a été constitué avec des quotas d'âge, c'est-à-dire qu'on a sélectionné des personnes au hasard jusqu'à atteindre un nombre fixé de personnes dans chaque

classe d'âge. Celles-ci ont été interrogées sur l'opérateur choisi. Le tableau ci-dessous donne la répartition des choix effectués en fonction de l'âge de l'utilisateur.

	opérateur 1	opérateur 2	opérateur 3
10-19	17	32	57
20-35	38	72	64
36-50	53	42	39
51 +	30	19	27

Identifier la situation d'échantillonnage appropriée et exprimer formellement l'hypothèse correspondant à la formulation suivante : il n'y a pas de relation entre l'âge et le type d'opérateur choisi. Tester cette hypothèse.

Exercice 10.12 Une enquête par sondage est menée parallèlement dans deux pays de l'Union Européenne sur la répartition des revenus dans une catégorie bien déterminée de salariés. On obtient les résultats suivants :

Salaire mensuel (euros)	Pays A	Pays B
<1200	4	6
1200-1600	22	18
1601-2000	20	18
>2000	14	6
Ensemble	60	48

Identifier la situation d'échantillonnage appropriée. La différence entre les répartitions des revenus observées dans les deux pays est-elle significative ?

Exercice 10.13 Les données du tableau qui suivent ont été étudiées par le statisticien belge A. Quetelet (1796-1874) et reprises de l'ouvrage de W.S. Peters : *Counting for Something* (Springer-Verlag, N.Y., 1986). Elles concernent les mesures (en pouces) de tour de poitrine de 5 738 soldats écossais.

Mesure	33	34	35	36	37	38	39	40
Fréquence	3	18	81	185	420	749	1073	1079
Mesure	41	42	43	44	45	46	47	48
Fréquence	934	658	370	92	50	21	4	1

Tester l'ajustement d'un modèle gaussien pour ces données.

Aide : il s'agit de données regroupées (par arrondi au pouce le plus proche) pour lesquelles on procédera comme indiqué en section 10.4.1 à ce propos.

Exercice 10.14 On a interrogé deux échantillons indépendants de 30 personnes chacun. Le premier échantillon est constitué de personnes appartenant à des ménages avec enfant(s), le deuxième de personnes appartenant à des ménages sans enfant. A la question «Considérez-vous que l'éducation des enfants est actuellement trop permissive ?» on a demandé aux enquêtés de

répondre en se positionnant sur une échelle ordonnée de 1 (oui, tout à fait) à 5 (non, pas du tout). Les résultats obtenus sont les suivants :

Réponse :	1	2	3	4	5	6
avec enfant	4	11	5	8	2	30
sans enfant	2	6	8	8	6	30
Ensemble	6	17	13	16	8	60

On veut tester l'hypothèse qu'il n'y a pas de différence d'attitude entre les deux types de personnes. Étant donné les tailles relativement réduites des échantillons d'une part, et le flou de la mesure effectuée d'autre part, on procédera à un test non paramétrique.

Aide : vu qu'il n'y a que 5 valeurs possibles on aura un nombre d'ex aequo important. On préférera donc le test de la médiane au test de Wilcoxon. Les observations correspondant à la valeur médiane générale devront être ignorées.

Chapitre 11

Régressions linéaire, logistique et non paramétrique

11.1 Introduction à la régression

A différentes reprises nous avons dit dans les chapitres précédents que tel ou tel résultat avait une validité au-delà du cadre strict des échantillons aléatoires constitués de variables aléatoires i.i.d.. L'objectif de ce chapitre est de montrer comment les méthodes classiques doivent être adaptées lorsque les v.a. observées restent indépendantes mais ne sont plus identiquement distribuées. Ceci peut être illustré avec profit dans les modèles explicatifs appelés modèles de régression.

De façon informelle, un modèle explicatif est un modèle exprimant une variable $\mathcal{Y}$, appelée *variable à expliquer* (ou réponse), comme une fonction d'une ou de plusieurs variables dites *variables explicatives* ou prédicteurs[1]. Toutefois si l'entité $\mathcal{Y}$ est considérée comme une variable aléatoire Y, un terme aléatoire, caractérisant l'incertitude de la prédiction, doit être introduit d'une certaine façon dans l'équation du modèle.

Dans *un modèle de régression*, on cherche essentiellement à déterminer la variation de **l'espérance mathématique** de Y en fonction des variables explicatives. En d'autres termes on étudie comment Y évolue «en moyenne» en fonction de ces variables explicatives. Dans ce chapitre, par souci de simplification, nous ne considérons qu'**une seule variable explicative** ce qui constitue la *régression simple* par opposition à la régression multiple. De plus cette entité

[1] Un tel modèle ne restituant pas nécessairement une relation de cause à effet directe le terme de prédicteur serait plus approprié.

explicative, que nous symboliserons par la lettre $\mathcal{X}$, sera une variable quantitative[2], pouvant prendre toute valeur dans un intervalle I de $\mathbb{R}$. Aux différentes valeurs de $\mathcal{X}$ dans I correspondent, par hypothèse, des v.a. distinctes et on est donc, en fait, en présence d'une famille de v.a. $\{Y(x) \mid x \in I\}$. Admettant que pour tout x l'espérance mathématique existe, alors $E(Y(x))$ est la fonction $g(x)$ qu'il s'agit de rechercher. Cette fonction mettant en évidence l'évolution moyenne de l'entité $\mathcal{Y}$ à expliquer en fonction de x est appelée *fonction de régression.* Dans cette approche on considère naturellement que l'incertitude de la prédiction de Y pour le «niveau» x de $\mathcal{X}$, se manifeste par une v.a. $\varepsilon(x)$ venant s'ajouter à la composante déterministe $g(x)$. Dans sa forme la plus générale un modèle de régression simple s'écrit donc :

$$Y(x) = g(x) + \varepsilon(x).$$

Puisque $E(Y(x)) = g(x)$, on a nécessairement $E(\varepsilon(x)) = 0$, quel que soit x. La v.a. $\varepsilon(x)$ est appelée *erreur* ou aléa (d'où la notation habituelle du «e» grec). Dans la plupart des modèles on suppose que l'erreur est de même loi quel que soit x ce qui permet d'écrire $Y(x) = g(x)+\varepsilon$ (on écrit même parfois simplement $Y = g(x)+\varepsilon$ en omettant d'indiquer que la v.a. Y est assujettie à la valeur x).

Le premier modèle que nous étudierons est le modèle de *régression linéaire* où $g(x) = \beta_0 + \beta_1 x$, que l'on écrira donc :

$$Y(x) = \beta_0 + \beta_1 x + \varepsilon.$$

Ce modèle est le plus simple qui soit et, de ce fait, est celui qui est utilisé le plus fréquemment. Il stipule qu'**en moyenne** l'entité $\mathcal{Y}$ varie linéairement en fonction du niveau de l'entité $\mathcal{X}$, ce qui est une hypothèse souvent réaliste. Par exemple le poids moyen des individus (adultes d'un même sexe) ayant une taille donnée x peut être considéré comme une fonction croissant linéairement avec x. La régression linéaire constitue le point de départ historique et méthodologique de toute la modélisation explicative. Ce modèle a été proposé par Francis Galton dans son ouvrage *Natural Inheritance* publié en 1889, notamment pour l'étude de la variation de la taille d'un homme en fonction de celle de son père. Il a choisi le terme de «régression» constatant qu'en moyenne un père grand tendra à avoir un fils plus petit que lui (et vice-versa pour un père petit).

Le deuxième modèle présenté dans le chapitre est le modèle logistique dont la particularité est que la variable à expliquer est binaire, du type «succès» ou «échec». On la codera comme précédemment par 1 ou 0 pour que $Y(x)$ soit, pour tout $x \in I$, une v.a. de Bernoulli. On essaie, par exemple, de déterminer dans quelle mesure le fait d'avoir ou de ne pas avoir d'incident cardiaque à un certain âge est lié au taux sanguin de cholestérol. Dans l'écriture de ce modèle de régression nous introduirons une fonction $g(x)$ particulièrement adaptée au fait que $Y(x)$ prend les valeurs 1 ou 0.

[2]La variable explicative pourrait être catégorielle ce qui, dans le cas de la régression linéaire, correspond à l'analyse de variance à un facteur.

Ces deux premiers types de modèles sont des *modèles paramétriques* car la fonction de régression $g(x)$ est de forme connue mais dépendant de paramètres inconnus (comme β_0 et β_1 dans le cas de la régression linéaire) qu'il s'agira d'estimer. Tous deux permettront d'illustrer l'application des méthodes paramétriques classiques des chapitres 6, 7 et 9. Dans le premier cas on obtient des solutions exactes simples. Dans le second cas on verra comment utiliser la méthode du maximum de vraisemblance et les propriétés asymptotiques de l'EMV lorsqu'il n'y a pas de solution explicite. L'intérêt de cette présentation réside dans le fait que la démarche est celle appliquée en statistique pour la plupart des modèles complexes.

Le troisième type de modèle sera le *modèle non paramétrique* où la fonction de régression $g(x)$ est totalement inconnue et doit être estimée. Nous sommes là face à un problème d'estimation fonctionnelle comme pour l'estimation d'une densité ou d'une fonction de répartition, vue en section 8.5.

Le modèle conditionnel

Dans notre exposé l'entité explicative $\mathcal{X}$ est une variable déterministe, ce qui a des implications sur la façon dont les observations sont effectuées. Cela suppose en effet que l'on se trouve dans des conditions expérimentales avec un **choix planifié des valeurs** $x_1, x_2, \cdots, x_n$ de $\mathcal{X}$, c'est-à-dire fixées à l'avance selon ce qu'on appelle un *plan d'expérience.* Supposons par exemple que l'on veuille étudier l'influence du taux d'engrais (en kg/hectare) sur le rendement (en tonnes/hectare) d'un type de céréale. On sème alors n parcelles expérimentales traitées avec des taux d'engrais choisis $x_1, x_2, \cdots, x_n$ (certaines valeurs pouvant être répétées). On considère que le rendement de chaque parcelle est une variable aléatoire du fait des multiples facteurs, autres que le taux d'engrais, qui le déterminent (pour preuve, les valeurs seront certainement différentes si l'on répète l'expérience). Les valeurs de rendement $y_1, y_2, \cdots, y_n$ effectivement observées seront donc traitées comme des réalisations de variables aléatoires $Y_1, Y_2, \cdots, Y_n$. Ici Y_i symbolise la loi du rendement pour un niveau d'engrais x_i, dont la moyenne est $E(Y_i) = g(x_i)$. Ces v.a. $Y_1, Y_2, \cdots, Y_n$ sont supposées indépendantes mais non de même loi puisqu'elles diffèrent au moins par leur moyenne. Il s'agira alors d'estimer, à partir de ces observations effectuées en quelques valeurs de $\mathcal{X}$ dans I, la fonction $g(x)$ pour tout $x \in I$, où I est la plage de variation du taux d'engrais qui intéresse l'expérimentateur.

Bien souvent, on ne se trouve pas dans de telles conditions expérimentales mais plutôt dans le cadre d'observations répétées d'un couple de v.a. (X, Y). Prenons, par exemple, le cas d'un sondage effectué pour étudier la variation du revenu $\mathcal{Y}$ en fonction de l'âge $\mathcal{X}$ dans une population. En tirant au hasard un individu on génère un couple aléatoire (X, Y) où X est la valeur de l'âge et Y est la valeur du revenu. Pour n individus on observe alors un échantillon aléatoire $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ à valeurs dans $\mathbb{R}^2$. Pour les valeurs effectivement observées $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, les $x_1, x_2, \cdots, x_n$ doivent être considérées comme des réalisations de v.a. au même titre que les

$y_1, y_2, \cdots, y_n$. En effet, contrairement aux circonstances précédentes les valeurs des x_i ne pouvaient être connues avant expérience, car résultant du processus de sélection au hasard. L'objectif essentiel restant d'étudier comment «en moyenne» le revenu varie en fonction de l'âge, la fonction de régression $g(x)$ est alors l'**espérance mathématique de la loi conditionnelle de Y sachant $X = x$**, notée $E(Y|X = x)$, soit :

$$g(x) = E(Y|X = x)\,.$$

Néanmoins, **dans le modèle classique de régression, les valeurs $x_1, x_2, \cdots, x_n$ ne sont pas traitées comme des réalisations de variables aléatoires,** ce qui facilite grandement les calculs. Stricto sensu les développements de ce chapitre ne sont valables que conditionnellement aux valeurs prises par les X_i. Le modèle de régression n'est donc, en principe, pas approprié dans une situation d'observations répétées d'un couple de variables aléatoires. Les propriétés d'optimalité conditionnelle des estimateurs, par exemple, peuvent être perdues lorsqu'on prend en compte le caractère aléatoire des X_i. Toutefois on constate généralement que ces estimateurs conservent des qualités assez proches, voire même identiques. Aussi le praticien applique-t-il les procédures de la régression quelles que soient les circonstances de la collecte des données.

Le formalisme conditionnel étant plus général car permettant d'envisager que la variable explicative puisse avoir un statut aléatoire, nous l'adopterons comme la plupart des auteurs. Mais il est clair que les résultats issus du modèle conditionnel s'appliqueront de la même façon à la situation décrite initialement avec l'exemple du rendement d'une céréale.

11.2 La régression linéaire

11.2.1 Le modèle

Nous supposons donc que la fonction de régression est de la forme :

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

pour x appartenant à un certain intervalle I. De plus nous supposons que la variance de la loi conditionnelle de Y sachant $X = x$ ne dépend pas de x et est égale à σ^2. Enfin nous faisons l'hypothèse que cette loi est gaussienne, quel que soit x. Nous reviendrons toutefois par la suite sur cette condition qui n'est pas cruciale. Notons que la linéarité du modèle est relative aux paramètres β_0 et β_1, et que l'on peut substituer à X des transformées $\ln X, \sqrt{X}, X^2$ etc., pour atteindre éventuellement la linéarité du modèle.

Le modèle contient donc trois paramètres inconnus : β_0, β_1 et σ^2. Pour estimer ces paramètres nous considérons une série d'observations indépendantes $Y_1, Y_2, \cdots, Y_n$ situées, respectivement, aux niveaux $x_1, x_2, \ldots, x_n$ de la variable explicative fixés (on suppose naturellement qu'au moins deux de ces valeurs sont

distinctes). Ainsi, pour tout i, $Y_i \rightsquigarrow \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Il est aussi commode d'utiliser la notation :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad i = 1, \ldots, n.$$

Pour tout i on a donc $\varepsilon_i \rightsquigarrow \mathcal{N}(0\,, \sigma^2)$. L'indépendance des Y_i entraîne celle des «erreurs» $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ qui sont donc des variables aléatoires i.i.d..

11.2.2 Les estimateurs du maximum de vraisemblance

La fonction de vraisemblance des trois paramètres, associée aux réalisations $y_1, y_2, \cdots, y_n$, est :

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}[y_i - (\beta_0 + \beta_1 x_i)]^2\}.$$

D'où la log-vraisemblance :

$$\ln L(\beta_0, \beta_1, \sigma^2) = -n(\ln\sqrt{2\pi} + \frac{1}{2}\ln\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2.$$

En annulant les dérivées partielles successivement par rapport β_0, β_1 et σ^2 on obtient les **équations de vraisemblance** :

$$\begin{cases} \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)] = 0 \\ \sum_{i=1}^{n} x_i[y_i - (\beta_0 + \beta_1 x_i)] = 0 \\ \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2 = 0\,. \end{cases}$$

Les deux premières équations ne dépendent pas de σ^2 et, étant linéaires en β_0 et β_1, peuvent être résolues. De la première équation on déduit $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$, puis en remplaçant dans la deuxième :

$$\widehat{\beta}_1(\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i) = \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i.$$

Or $\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$ selon la formule bien connue de centrage-décentrage de la statistique descriptive et, de la même façon, $\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$. D'où finalement, en substituant les Y_i aux y_i, les estimateurs du MV de $\widehat{\beta}_0$ et $\widehat{\beta}_1$:

$$\begin{cases} \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x} \\ \widehat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{cases} .$$

Il est intéressant de noter que les deux premières équations de vraisemblance correspondent à la minimisation du terme $\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$ dans l'expression de la log-vraisemblance. Pour une solution quelconque (β_0^*, β_1^*), la

différence $y_i - (\beta_0^* + \beta_1^* x_i)$ est appelée *résidu* car elle correspond à l'écart entre la valeur observée et celle donnée par le modèle ainsi estimé. On voit donc que le couple $(\widehat{\beta}_0, \widehat{\beta}_1)$ est la solution qui minimise la *somme des carrés des résidus.* De ce fait $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont aussi appelés *estimateurs des moindres carrés.*

Montrons que $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont respectivement sans biais pour β_0 et β_1. On a :

$$E(\overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1 \overline{x}$$
$$E(Y_i - \overline{Y}) = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \overline{x}) = \beta_1 (x_i - \overline{x}).$$

D'où :

$$E(\widehat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})E(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta_1.$$
$$E(\widehat{\beta}_0) = E(\overline{Y}) - E(\widehat{\beta}_1)\overline{x} = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0.$$

On en déduit que $\widehat{\beta}_0 + \widehat{\beta}_1 x$ est sans biais pour $E(Y|X = x)$, l'espérance de la réponse pour la valeur x de la variable explicative.

Pour calculer les variances de ces estimateurs notons que le numérateur de l'expression de $\widehat{\beta}_1$ s'écrit aussi $\sum_{i=1}^{n}(x_i - \overline{x})Y_i$ puisque $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. Donc :

$$V(\widehat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 V(Y_i)}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Pour $\widehat{\beta}_0$ on a $V(\widehat{\beta}_0) = V(\overline{Y}) + \overline{x}^2 V(\widehat{\beta}_1) - 2\,\overline{x}\, cov(\overline{Y}, \widehat{\beta}_1)$. Or :

$$cov(\overline{Y}, \widehat{\beta}_1) = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} cov\left(\overline{Y}\,,\, \sum_{i=1}^{n}(x_i - \overline{x})Y_i\right)$$
$$= \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \sum_{i=1}^{n}(x_i - \overline{x}) cov\left(\overline{Y}\,,\, Y_i\right)$$

et, comme $cov(Y_j, Y_i) = 0$ si $j \neq i$, $cov\left(\overline{Y}\,,\, Y_i\right) = cov\left(\frac{1}{n}Y_i\,,\, Y_i\right) = \frac{\sigma^2}{n}$, on a :

$$cov(\overline{Y}, \widehat{\beta}_1) = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} \sum_{i=1}^{n}(x_i - \overline{x}) = 0\,,$$

d'où :

$$V(\widehat{\beta}_0) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right).$$

De plus :

$$cov(\widehat{\beta}_0, \widehat{\beta}_1) = cov(\overline{Y} - \widehat{\beta}_1 \overline{x}\,,\, \widehat{\beta}_1) = cov(\overline{Y}, \widehat{\beta}_1) - \overline{x}\, V(\widehat{\beta}_1) = -\frac{\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

En posant $\boldsymbol{\beta} = (\beta_0, \beta_1)^t$ et $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)^t$ nous résumons ces résultats (avec les notations pour les vecteurs aléatoires introduites en section 3.8) par :

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \quad \text{et} \quad \mathbb{V}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} \sigma^2 \left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) & -\dfrac{\overline{x}\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ -\dfrac{\overline{x}\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} & \dfrac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \end{pmatrix}.$$

Comme $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont chacun une combinaison linéaire des Y_i, d'après le théorème de caractérisation 3.1 le vecteur aléatoire $\widehat{\boldsymbol{\beta}}$ est gaussien. Sa loi est donc parfaitement définie.

Déterminons maintenant l'estimateur du MV de σ^2. Il se déduit de la dernière équation de vraisemblance selon :

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2 \,,$$

expression dans laquelle on retrouve la somme des carrés des résidus (sous sa forme aléatoire). Pour étudier la distribution d'échantillonnage de $\widehat{\sigma}^2$ on admettra la proposition suivante.

Proposition 11.1 *La v.a.* $\frac{1}{\sigma^2} \sum_{i=1}^n [Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2$ *suit une loi du khi-deux à* $n-2$ *degrés de liberté. De plus elle est indépendante de l'estimateur* $\widehat{\boldsymbol{\beta}}$.

Pour la première assertion cette proposition est à rapprocher du théorème 5.1 concernant la variance d'un échantillon, la démonstration faisant appel à des considérations similaires. La perte de deux degrés de liberté s'explique par le fait qu'il y a deux liaisons linéaires déterministes entre les $Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$ correspondant aux deux premières équations de vraisemblance. L'indépendance est à rapprocher de celle vue en proposition 5.3 entre moyenne et variance empiriques.

De cette proposition nous déduisons que $E(\sum_{i=1}^n [Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2) = (n-2)\sigma^2$. L'espérance mathématique de $\widehat{\sigma}^2$ est donc $\frac{n-2}{n}\sigma^2$: cet estimateur est biaisé. C'est pourquoi on lui préfère l'estimateur sans biais, obtenu en divisant la somme des carrés des résidus par $n-2$, que nous noterons S^2, soit :

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n [Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2.$$

On montre que les estimateurs $\widehat{\beta}_0$, $\widehat{\beta}_1$ et S^2 sont chacun UMVUE.

Si les x_i sont choisis de telle sorte que pour tout n leur variance descriptive $\sum_{i=1}^n (x_i - \overline{x})^2/n$ admette une borne inférieure strictement positive indépendante de n et leur moyenne $\overline{x}$ admette une borne supérieure également indépendante de n, on voit immédiatement que $V(\widehat{\beta}_0)$ et $V(\widehat{\beta}_1)$ tendent vers 0 quand $n \to \infty$. Alors $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont convergents en moyenne quadratique. Sachant que la variance d'une v.a. de loi $\chi^2(n-2)$ est $2(n-2)$ on a $V(S^2) = 2\sigma^4/(n-2)$ et S^2 converge aussi (vers σ^2) en moyenne quadratique.

11.2.3 Intervalles de confiance

Pour β_0 et β_1 on dispose d'une quantité pivot de même type que celle utilisée pour la moyenne d'une loi de Gauss en section 7.4.1. Pour β_1, par exemple, on a :

$$\frac{\widehat{\beta}_1 - \beta_1}{\sigma / \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \rightsquigarrow \mathcal{N}(0\,;1)$$

d'où, en estimant σ^2 par S^2, on obtient comme pour le théorème 5.2 :

$$\frac{\widehat{\beta}_1 - \beta_1}{S / \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \rightsquigarrow t(n-2).$$

On posera $S^2(\widehat{\beta}_1) = S^2 / \sum_{i=1}^n (x_i - \overline{x})^2$ pour noter l'estimateur de la variance de $\widehat{\beta}_1$. La variable aléatoire $(\widehat{\beta}_1 - \beta_1)/S(\widehat{\beta}_1)$ est donc une fonction pivot qui conduit immédiatement à l'intervalle de confiance suivant :

$$IC_{0,95}(\beta_1) = [\widehat{\beta}_1 - t_{0,975}^{(n-2)}\, s(\widehat{\beta}_1)\,;\, \widehat{\beta}_1 + t_{0,975}^{(n-2)}\, s(\widehat{\beta}_1)]$$

où :

$$s(\widehat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\,,$$

s désignant la réalisation de S (et $\widehat{\beta}_1$ désignant indifféremment l'estimation ou l'estimateur de β_1). De même on obtient :

$$IC_{0,95}(\beta_0) = [\widehat{\beta}_0 - t_{0,975}^{(n-2)}\, s(\widehat{\beta}_0)\,;\, \widehat{\beta}_0 + t_{0,975}^{(n-2)}\, s(\widehat{\beta}_0)]$$

où $s(\widehat{\beta}_0) = s\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}$.

Il est intéressant de construire un IC pour l'espérance de la réponse $\beta_0 + \beta_1 x$ au niveau x de la variable explicative. On a :

$$\begin{aligned} V(\widehat{\beta}_0 + \widehat{\beta}_1 x) &= V(\widehat{\beta}_0) + 2x\, cov(\widehat{\beta}_0, \widehat{\beta}_1) + x^2\, V(\widehat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} - \frac{2x\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{x^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right). \end{aligned}$$

Par des développements tout à fait analogues aux précédents on obtient :

$$IC_{0,95}(\beta_0 + \beta_1 x) = \\ [\widehat{\beta}_0 + \widehat{\beta}_1 x - t_{0,975}^{(n-2)}\, s(\widehat{\beta}_0 + \widehat{\beta}_1 x)\,; \widehat{\beta}_0 + \widehat{\beta}_1 x + t_{0,975}^{(n-2)}\, s(\widehat{\beta}_0 + \widehat{\beta}_1 x)]$$

où :

$$s(\widehat{\beta}_0 + \widehat{\beta}_1 x) = s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}.$$

On constate que la largeur de l'IC est d'autant plus grande que l'on s'éloigne de la valeur centrale $\overline{x}$ des valeurs fixées pour la variable explicative.

On peut également établir (voir les exercices) un *intervalle de prédiction* pour **une** observation au niveau x de la variable explicative.

11.2.4 Test $H_0 : \beta_1 = 0$

Ce test est essentiel car il décide de l'intérêt du modèle (ou de la «significativité» de la variable explicative). En utilisant le résultat de la section précédente on déduit que, sous H_0, $\widehat{\beta}_1 / S(\widehat{\beta}_1)$ suit une loi de Student à $n-2$ degrés de liberté. On rejettera donc H_0 au niveau α si :

$$\frac{\widehat{\beta}_1}{s(\widehat{\beta}_1)} \notin [-t_{1-\alpha/2}^{(n-2)}\,,\, t_{1-\alpha/2}^{(n-2)}].$$

Ce test est uniformément plus puissant parmi les tests sans biais. Comme pour le test de Student usuel vu en section 9.7.1 sa puissance pour une vraie valeur β_1 se lit sur une loi de Student non centrale de paramètre de non centralité $\beta_1 / \sqrt{V(\widehat{\beta}_1)}$. On ne peut obtenir qu'une valeur approchée de cette puissance du fait que $V(\widehat{\beta}_1)$ doit être estimé.

Notons que si le test est accepté la loi conditionnelle de Y sachant $X = x$ ne dépend pas de x : les v.a. Y_i sont toutes i.i.d. de loi $\mathcal{N}(\beta_0, \sigma^2)$ et l'on se retrouve dans la situation classique d'un échantillon issu d'une loi de Gauss.

Approche par l'analyse de variance[3]

On peut préférer aborder ce test par la voie de la relation de *décomposition de la somme des carrés totale*. En effet cette relation et la formulation du test qui s'ensuit ont une validité générale pour tout type de modèle linéaire (en

[3]Ne pas confondre avec le modèle d'analyse de variance qui concerne des variables explicatives catégorielles.

particulier, on indiquera cela pour la régression multiple en section 11.2.7). Posons, pour simplifier les développements, $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. La relation est :

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2.$$

Sa démonstration est proposée dans les exercices. Le premier terme est appelé *somme des carrés totale* car il exprime la variabilité des y_i indépendamment de tout modèle explicatif. Le deuxième terme se nomme *somme des carrés expliquée par le modèle* du fait qu'il ne prend en compte que les valeurs modélisées dont il rend compte de la variabilité (on vérifiera sans peine que la moyenne des $\widehat{y}_i$ est égale à $\overline{y}$). Le troisième terme est la somme des carrés des résidus. Démontrons la proposition suivante concernant la somme des carrés expliqués en tant que variable aléatoire : $\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$.

Proposition 11.2 *Sous l'hypothèse* $H_0 : \beta_1 = 0$, *on a :*

$$\frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sigma^2} \rightsquigarrow \chi^2(1).$$

Démonstration : en vertu de la relation de décomposition exprimée en termes aléatoires on peut écrire :

$$\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2\,,$$

soit, en substituant $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}$:

$$\begin{aligned}\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}\left[(Y_i - \overline{Y}) - \widehat{\beta}_1(x_i - \overline{x})\right]^2 \\ &= 2\widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) - \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2\,.\end{aligned}$$

De l'expression de $\widehat{\beta}_1$ en section 11.2.2 on déduit :

$$\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \widehat{\beta}_1 \sum_{i=1}^{n}(x_i - \overline{x})^2$$

d'où, en substituant ce terme :

$$\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \widehat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2\,.$$

Or, sous l'hypothèse $\beta_1 = 0$, $\widehat{\beta}_1 \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/\sigma^2}$ suit une loi $\mathcal{N}(0\,;1)$ et son carré suit une loi $\chi^2(1)$, ce qui prouve la proposition. □

Selon la proposition 11.1, $\frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$ est indépendant de $\widehat{\beta}_1$ et donc de $\frac{1}{\sigma^2}\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2 = \frac{1}{\sigma^2}\widehat{\beta}_1^2 \sum_{i=1}^n (x_i - \overline{x})^2$. Sous H_0 ces deux v.a. suivent, respectivement, des lois $\chi^2(n-2)$ et $\chi^2(1)$. Le rapport de la seconde à la première, après division par leurs degrés de liberté, suit donc une loi de Fisher $F(1, n-2)$. Le paramètre σ^2 disparaissant et la somme des carrés des résidus s'écrivant $(n-2)S^2$, ce rapport est :

$$F = \frac{\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2}{S^2}.$$

Or, de façon générale, $E(S^2) = \sigma^2$ et :

$$\begin{aligned} E\Big(\sum_{i=1}^n (\widehat{Y}_i - \overline{Y})^2\Big) &= E(\widehat{\beta}_1^2) \sum_{i=1}^n (x_i - \overline{x})^2 \\ &= \left[V(\widehat{\beta}_1) + \Big\{ E(\widehat{\beta}_1) \Big\}^2 \right] \sum_{i=1}^n (x_i - \overline{x})^2 \\ &= \sigma^2 + \beta_1^2 \sum_{i=1}^n (x_i - \overline{x})^2. \end{aligned}$$

On voit que si $\beta_1 \neq 0$ la somme des carrés expliquée tendra à être supérieure à S^2 et F peut être envisagée comme statistique de test incitant à rejeter H_0 pour des **valeurs trop élevées** sur la loi $F(1, n-2)$. En effet il n'y a pas lieu de rejeter H_0 lorsque F est faible puisque cela abonde dans le sens de l'hypothèse $\beta_1 = 0$. Notons que ce test est équivalent au test de Student présenté initialement car, d'une part, la statistique F est le carré de la statistique $\widehat{\beta}_1 / S(\widehat{\beta}_1)$ et, d'autre part (voir en fin de section 5.5), le carré d'une v.a. de loi $t(n-2)$ est une v.a. de loi $F(1, n-2)$.

L'usage veut que l'on présente les résultats menant à cette statistique sous forme d'un *tableau d'analyse de variance* (voir le tableau de l'exemple 11.1 ci-après), cette appellation venant du fait que l'on y met en évidence les termes de la décomposition de la variabilité totale à travers les sommes de carrés.

11.2.5 Cas non gaussien

Le couple $(\widehat{\beta}_0, \widehat{\beta}_1)$ n'est plus estimateur du MV mais demeure l'estimateur des moindres carrés de (β_0, β_1) car minimisant $\sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$. Par analogie avec l'estimation de la moyenne d'une loi à partir de la moyenne empirique $\overline{x}$ qui est la valeur de a minimisant $\sum_{i=1}^n (x_i - a)^2$, l'estimation de la droite de régression $y = \beta_0 + \beta_1 x$ par la droite des moindres carrés $y = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ semble être assez naturelle. Ceci est corroboré par le théorème suivant qui s'applique à toute la modélisation linéaire et y justifie la prééminence de la méthode des moindres carrés.

Théorème 11.1 (Gauss-Markov) *Les estimateurs des moindres carrés* $\widehat{\beta}_0$ *et* $\widehat{\beta}_1$ *sont, respectivement, estimateurs de variance minimale pour* β_0 *et* β_1 *parmi les estimateurs sans biais fonctions* ***linéaires*** *des* Y_i.

La démonstration est proposée dans la section des exercices. En réalité cette proposition s'étend à toute combinaison linéaire de $\widehat{\beta}_0$ et $\widehat{\beta}_1$, et en particulier, pour tout x, $\widehat{\beta}_0 + \widehat{\beta}_1 x$ est estimateur de variance minimale de $\beta_0 + \beta_1 x$ parmi les estimateurs linéaires sans biais. Bien que la classe où $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont optimaux soit plus réduite qu'avec l'hypothèse gaussienne, la propriété n'en reste pas moins intéressante, d'autant plus qu'elle n'exige aucune hypothèse sur le type de loi des erreurs (hormis bien sûr l'existence de la variance qui est implicite). Notons que les trois estimateurs $\widehat{\beta}_0, \widehat{\beta}_1$ et S^2 restent sans biais car les démonstrations précédentes ne recouraient pas à la nature gaussienne des Y_i. Pour $\widehat{\beta}_0$ et $\widehat{\beta}_1$ la convergence en moyenne quadratique est assurée dans les mêmes conditions que ci-dessus. Quant à celle de S^2 elle n'exige que l'existence du moment d'ordre 4 de la loi des erreurs.

Les tests et intervalles de confiance sont étonnamment robustes, le théorème central limite agissant indirectement. Toutefois cette robustesse connaît des limitations de même nature que pour l'inférence sur les moyennes de lois. En premier lieu la loi ne doit pas produire de valeurs extrêmes (i.e. pas de queues de distribution trop allongées). Par ailleurs la condition de variance constante («homoscédasticité») ne peut être assouplie que dans une faible mesure. On peut éventuellement recourir à une transformation de Y pour stabiliser la variance (par exemple par la fonction Arcsin si Y est une proportion, par sa racine carrée si c'est un comptage de type Poisson).

11.2.6 Régression et corrélation linéaires

Considérons un couple de v.a. (X, Y) et adoptons les notations μ_X, μ_Y, σ_X^2, σ_Y^2 et ρ pour leurs moyennes, leurs variances et leur coefficient de corrélation linéaire. Supposons que la fonction de régression $E(Y|X = x)$ soit linéaire. On montre alors (voir exercices) qu'elle est nécessairement de la forme :

$$E(Y|X = x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X).$$

Donc on a, dans les notations précédentes, $\beta_1 = \rho \frac{\sigma_Y}{\sigma_X}$. On peut constater immédiatement que la même relation vaut pour les estimateurs correspondants, i.e. $\widehat{\beta}_1 = R \frac{S_Y}{S_X}$ où R est la corrélation linéaire empirique définie à la fin de la section 5.2. On en déduit que R est l'estimateur du maximum de vraisemblance de ρ. En effet :

$$R = \widehat{\beta}_1 \frac{S_X}{S_Y} = \widehat{\beta}_1 \frac{\widetilde{S}_X}{\widetilde{S}_Y}$$

et $\widehat{\beta}_1$, $\widetilde{S}_X$, $\widetilde{S}_Y$ sont les estimateurs du MV respectifs de β_1, σ_X, σ_Y ($\widetilde{S}_X^2$ et $\widetilde{S}_Y^2$ désignent les variances empiriques, S_X^2 et S_Y^2 les variances d'échantillon, voir définitions 5.4 et 5.5).

Comme σ_Y et σ_X sont strictement positifs, $\beta_1 = 0$ si et seulement si $\rho = 0$. En particulier les hypothèses $H_0 : \beta_1 = 0$ et $H_0' : \rho = 0$ sont équivalentes. Par conséquent, si l'on dispose d'un test pour l'une des hypothèses il vaut pour l'autre. Nous pouvons appliquer cela au cas d'un vecteur gaussien car il est facile de montrer que la fonction de régression est linéaire (voir exercices). Distinguons les deux situations de recueil des données, à savoir avec un plan d'expérience (où les valeurs de X sont fixées a priori) ou avec des observations de X elles-mêmes aléatoires. Dans le premier cas le test $H_0 : \beta_1 = 0$ de la section 11.2.4 peut être considéré également comme un test de non corrélation entre les deux variables aléatoires X et Y. Nous examinons maintenant plus en détail la deuxième situation.

Soit un échantillon de taille $n : (X_1,Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$. Nous avons vu en section 9.7.7 un test de $H_0 : \rho = 0$ fondé sur un tel échantillon. Ce test reposait sur la statistique $\sqrt{n-2}R/\sqrt{1-R^2}$ qui, sous H_0, suit une loi $t(n-2)$. Nous avons admis ce résultat que nous sommes maintenant en mesure de démontrer. Notons $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ et $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ une réalisation de $\mathbf{X}$. Considérons la statistique F vue plus haut (section 11.2.4) dans laquelle on remplace les x_i par les v.a. X_i. Les développements précédents indiquent que la loi de F **conditionnellement** à $\mathbf{X} = \mathbf{x}$ est, sous H_0, la loi $F(1, n-2)$. Cette loi ne dépend pas de $\mathbf{x}$ ce qui signifie que, sous H_0, la statistique F est indépendante de $\mathbf{X}$. La loi non conditionnelle de F (ou loi marginale) est donc aussi la loi $F(1, n-2)$. Ainsi F calculé à partir des couples (X_i, Y_i) peut être utilisé comme statistique de test pour $H_0 : \rho = 0$. Pour ce qui concerne la mise en oeuvre ce test ne se distingue donc pas de celui vu plus haut lorsque les x_i sont fixés. Montrons qu'on a affaire au même test que celui proposé en section 9.7.7 et, pour simplifier, raisonnons sur les réalisations.

D'une relation établie lors de la démonstration de la proposition 11.2, on déduit :

$$\begin{aligned} r^2 &= \frac{\left[\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right]^2}{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2} \\ &= \frac{\widehat{\beta}_1^2 \sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} \\ &= \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}. \end{aligned}$$

En raison de la décomposition de la somme des carrés totale on voit que le rapport de la somme des carrés des résidus à cette dernière est égal à $1 - r^2$ (r^2 est appelé coefficient de détermination, voir section 9.7.7). Donc F (réalisé) peut s'écrire :

$$F = \frac{(n-2)r^2}{1-r^2}.$$

Sa racine carrée est la réalisation de la statistique de Student (en valeur absolue) de la section 9.7.7. Or une v.a. de loi $t(n-2)$ élevée au carré est une v.a. de loi $F(1, n-2)$ ce qui prouve que les deux tests sont identiques.

Il est intéressant de noter que la loi non conditionnelle de F est une loi $F(1, n-2)$, sous H_0, quelle que soit la loi marginale de X. Pour que le test s'applique rigoureusement il suffit donc que, pour tout x, la loi conditionnelle de Y sachant $X = x$ soit gaussienne de variance σ^2 indépendante de x et que la fonction de régression de Y sur X soit linéaire.

Nous venons de voir que le test usuel $H_0 : \beta_1 = 0$ est utilisable même si les x_i sont des réalisations de v.a. X_i. On montre que les autres résultats établis conditionnellement aux x_i fixés restent également valables à condition que la loi marginale de X ne dépende pas des paramètres β_0 et β_1 définissant la fonction de régression de Y sur X.

Exemple 11.1 Pour une enquête on a eu recours à 54 enquêteurs. Pour chacun d'entre eux on dispose du nombre d'entretiens qu'il a effectués et de la durée médiane de ceux-ci[4]. On cherche à vérifier si le nombre d'entretiens effectués X est un facteur explicatif de la durée de l'entretien Y. On a calculé initialement :

$$\overline{x} = 53 \; ; \; \overline{y} = 30{,}535 \; ; \; \sum_{i=1}^{54} x_i^2 = 4274{,}8 \; ; \; \sum_{i=1}^{54} y_i^2 = 957{,}23 \; ; \; \sum_{i=1}^{54} x_i y_i = 1531{,}7 \, .$$

On en déduit les estimations suivantes :

$$\widehat{\beta}_0 = 33{,}668 \quad ; \quad \widehat{\beta}_1 = -0{,}05911 \quad ; \quad s^2 = 20{,}473$$
$$s^2(\widehat{\beta}_0) = 1{,}1057 \quad ; \quad s^2(\widehat{\beta}_1) = 0{,}0002589 \quad ; \quad \widehat{cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -0{,}01371.$$

Pour le test de l'hypothèse $H_0 : \beta_1 = 0$, la statistique de test prend la valeur $t = \widehat{\beta}_1 / \sqrt{s^2(\widehat{\beta}_1)}$= -3,68 ce qui correspond, pour la loi de Student $t(52)$, à une P-valeur de l'ordre de 0,001. Le nombre d'entretiens effectués est donc un facteur explicatif très significatif de la durée médiane de ces entretiens. Le même test peut être conduit à partir du tableau d'analyse de variance ci-après.

Source	Somme des Carrés	ddl	Carrés Moyens	F	P-valeur
Expliquée	276,53	1	276,53	13,51	0,001
Résiduelle	1 064,59	52	20,47		
Totale	1 341,12	53			

Pour un niveau $x = 50$ entretiens on obtient une estimation de l'espérance de la durée médiane des entretiens égale à :

$$\widehat{\beta}_0 + \widehat{\beta}_1 . 50 = 30{,}713$$

[4]Source : Centre d'Etudes des Supports de Publicité, Paris.

et un intervalle de confiance de niveau 0,95 associé :

$$IC_{0,95}(\beta_0 + \beta_1.50) = [30{,}713 - t^{(52)}_{0,975}.0{,}618\ ; 30{,}713 + t^{(52)}_{0,975}.0{,}618]$$

soit, avec $t^{(52)}_{0,975} = 2{,}007$, l'intervalle [29,47 ; 31,95]. La figure 11.1 indique les limites de confiance en fonction du niveau de x ainsi que les limites de prédiction (plus larges) décrites dans les exercices. Au vu de la dispersion des points autour de la droite de régression la condition de variance constante est plausible. ■

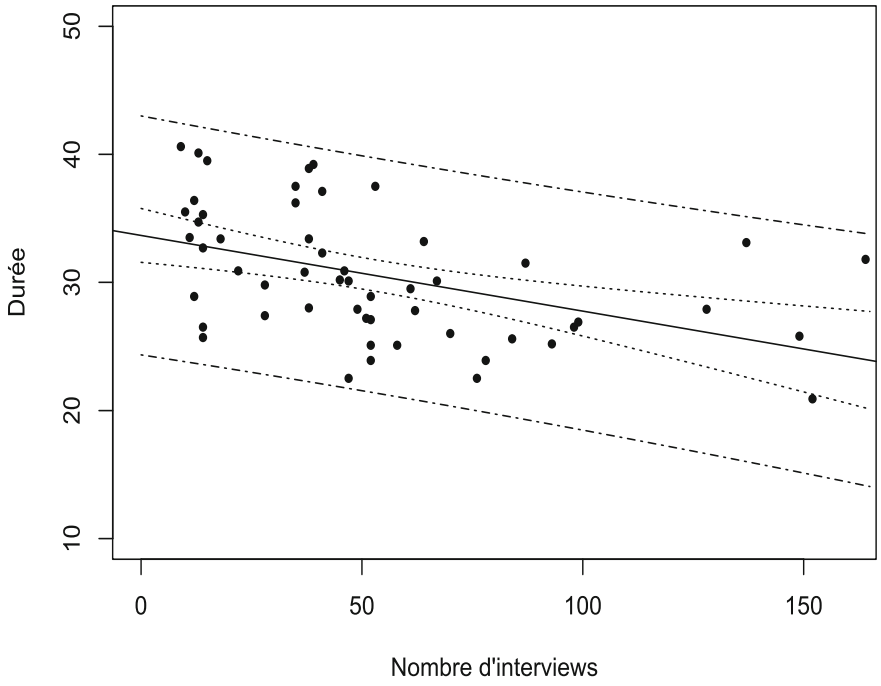


Figure 11.1 - Données «enquêteurs» : droite de régression, limites de confiance, limites de prédiction.

11.2.7 Extension à la régression multiple

Dans cette section nous montrons succinctement que l'étude du modèle de régression simple s'étend sans difficultés en présence de plusieurs variables explicatives. Dans le formalisme du modèle conditionnel présenté en section 11.1 on considère p prédicteurs $X_1, X_2, \ldots, X_p$ et une fonction de régression de

Y de la forme :

$$E(Y \mid X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

pour les niveaux respectifs $x_1, x_2, \ldots, x_p$ de ces prédicteurs. Les autres hypothèses restent identiques : lois conditionnelles de Y gaussiennes et de même variance σ^2. Soit une série d'observations indépendantes $Y_1, Y_2, \ldots, Y_n$ où Y_i est observé pour les valeurs $x_{i1}, x_{i2}, \ldots, x_{ip}$ des variables explicatives. On recourt alors à l'écriture matricielle suivante. On définit le vecteur des observations $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^t$, puis le vecteur des $p+1$ paramètres inconnus $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^t$ et la $n \times (p+1)$-matrice du plan d'expérience $\mathbf{X}$ dont la i-ème ligne est $(1, x_{i1}, x_{i2}, \ldots, x_{ip})$. On a alors, avec les notations de la section 3.8 :

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathbb{V}(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$$

où $\mathbf{I}_n$ désigne la matrice identité d'ordre n.

La log-vraisemblance s'écrit comme en section 11.2.2 en remplaçant $\beta_0 + \beta_1 x_i$ par $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. Les équations de vraisemblance obtenues en dérivant par rapport à chacun des $p+1$ paramètres forment un système linéaire de $p+1$ équations dont l'écriture matricielle est $(\mathbf{X}^t\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$. On suppose que les vecteurs colonnes de $\mathbf{X}$ sont linéairement indépendants (i.e. $n > p$ et pas de redondance d'information dans les vecteurs prédicteurs ni de combinaison linéaire entre eux donnant un vecteur constant, ce qui signifierait une surparamétrisation du modèle). Alors la matrice $\mathbf{X}^t\mathbf{X}$ est inversible et on a la solution unique :

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}\,.$$

$\widehat{\boldsymbol{\beta}}$ est également le vecteur des estimateurs des moindres carrés, c'est-à-dire tel que :

$$\left\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|^2 = \min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

où $\|.\|^2$ représente la norme euclidienne usuelle d'un vecteur de $\mathbb{R}^n$. D'après la proposition 3.13 on a :

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\mathbb{E}(\mathbf{Y}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$
$$\mathbb{V}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\,\mathbb{V}(\mathbf{Y})\,\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}.$$

Sachant que $\widehat{\boldsymbol{\beta}} \rightsquigarrow \mathcal{N}_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1})$ on peut procéder à tout type d'inférence concernant les paramètres du modèle. Il est notamment intéressant de tester des hypothèses du type $H_0 : \beta_k = 0$ permettant (pour $k \geq 1$) de décider de la pertinence de tel prédicteur particulier en présence des autres prédicteurs. La statistique de test est analogue à celle de la régression simple pour

$H_0 : \beta_1 = 0$ à cette différence près que le nombre de degrés de liberté de la loi de Student devient égal à $n - (p + 1)$. Ceci découle du fait que la somme des carrés des résidus ne comporte plus que $n - (p + 1)$ degrés de liberté. Compte tenu de cette modification on peut construire un tableau d'analyse de variance semblable à celui de la régression simple. La statistique F permet alors de tester l'hypothèse globale $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, les valeurs critiques se rapportant à la loi $F(p, n - (p + 1))$.

Au-delà du modèle de régression qui stipule l'existence d'une fonction de régression, un *modèle linéaire* dans sa forme la plus générale se définit comme un vecteur d'observations $\mathbf{Y}$ tel que $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ et $\mathbb{V}(\mathbf{Y}) = \sigma^2\mathbf{I}_n$ comme ci-dessus. Ceci inclut notamment les modèles d'analyse de variance où les variables explicatives (appelées alors facteurs) sont catégorielles ce qui conduit à introduire dans la matrice $\mathbf{X}$ des variables indicatrices des différentes catégories induites par ces facteurs.

Les *modèles linéaires généralisés* constituent un vaste ensemble d'extensions du modèle de régression multiple où, d'une part, les Y_i répondent à d'autres types de lois paramétriques que la loi de Gauss et, d'autre part, la fonction de régression s'exprime sous la forme $g(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$ où g est une fonction connue. Le modèle logistique présenté ci-après en offre une illustration particulièrement importante.

Pour un traitement plus complet de la régression linéaire on pourra consulter, pour les aspects mathématiques, l'ouvrage classique de Seber (1977) et, pour les aspects pratiques, le livre de Dodge (1999).

11.3 La régression logistique

11.3.1 Le modèle

Ce modèle est adapté au cas où la variable à expliquer est binaire. En utilisant le codage 1/0 on la transforme en variable aléatoire de Bernoulli. Plus précisément, dans le formalisme conditionnel exposé en section 11.1, la loi de Y sachant $X = x$ est une loi $\mathcal{B}(p(x))$. La fonction de régression à estimer est donc :

$$E(Y|X = x) = p(x)$$

où $p(x) = P(Y = 1|X = x)$. Plus prosaïquement, le problème est de déterminer comment la probabilité de «succès» évolue en fonction du niveau de la variable X. Par exemple : quelle est la probabilité que le client d'une banque détienne des valeurs mobilières, en fonction de son niveau de revenu ?

Nous ne sommes donc plus dans le cadre précédent et le modèle de régression linéaire usuel n'est, en principe, pas approprié. Nous disons «en principe» car ce modèle est très robuste vis-à-vis de l'hypothèse gaussienne dans la mesure où

l'on a un nombre suffisant d'observations, au même titre que l'approximation d'une loi binomiale par une loi de Gauss. Mais le modèle linéaire pose problème pour une raison majeure : la fonction $p(x)$ n'y est pas contrainte dans l'intervalle $[0,1]$ et les estimations peuvent donc produire des valeurs négatives ou supérieures à 1. Le modèle logistique remédie à cela.

Ce modèle stipule que la probabilité conditionnelle de succès est de la forme :

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\,.$$

La fonction :

$$g(u) = \frac{e^u}{1+e^u}$$

est appelée *fonction logistique,* elle est strictement croissante et prend ses valeurs dans l'intervalle [0,1] (voir figure 11.2). Sa fonction inverse est :

$$g^{-1}(u) = \ln \frac{u}{1-u}$$

et s'appelle *fonction logit*[5]. Pour une loi de Bernoulli $\mathcal{B}(p)$ le rapport $\frac{p}{1-p}$ a une certaine signification. On l'appelle parfois la *chance* ou la cote de succès (en anglais : *odds*). Dans le modèle logistique le logarithme de ce rapport est donc une fonction linéaire de la variable explicative :

$$\ln \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x.$$

Le modèle comporte donc deux paramètres inconnus β_0 et β_1. On notera par $\boldsymbol{\beta}$ le couple (β_0,β_1) ou, indifféremment, le vecteur $(\beta_0,\beta_1)^t$. Contrairement à la régression classique il n'y a pas de variance de l'erreur à estimer puisqu'une loi de Bernoulli $B(p(x))$ ne dépend que du paramètre $p(x)$.

11.3.2 Estimation de la fonction $p(x)$

Supposons que nous observions indépendamment les v.a. binaires $Y_1, Y_2, \cdots, Y_n$ aux points $x_1, x_2, \cdots, x_n$ de la variable explicative et déterminons l'estimateur du maximum de vraisemblance de $\boldsymbol{\beta}$. Pour tout i, $Y_i \rightsquigarrow B(p(x_i))$ et la fonction de probabilité de Y_i est (voir section 4.1.2) :

$$p(y) = p(x_i)^y (1-p(x_i))^{1-y}\,,\ y \in \{0,1\}\,.$$

La fonction de vraisemblance de $\boldsymbol{\beta}$ associée à une réalisation $(y_1, y_2, \cdots, y_n)$ de $(Y_1, Y_2, \cdots, Y_n)$ est donc :

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(x_i)^{y_i} (1-p(x_i))^{1-y_i}$$

[5]Ceci amène une confusion entre modèle logit et modèle logistique. L'usage le plus répandu est de parler de régression logistique lorsque, comme ici, la (ou les) variable explicative est quantitative et de modèle logit lorsqu'elle (ou elles) est catégorielle.

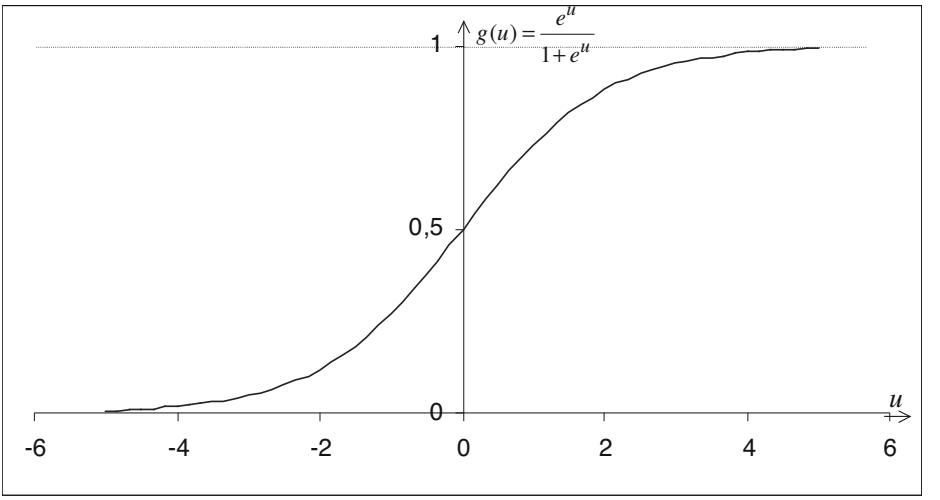


Figure 11.2 - Fonction logistique.

avec :

$$p(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} .$$

La log-vraisemblance est égale à :

$$\ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \ln p(x_i) + (1 - y_i) \ln[1 - p(x_i)]\} .$$

Les deux équations de vraisemblance sont établies en dérivant cette fonction par rapport à β_0 et par rapport à β_1. Dans un premier temps considérons la dérivée de la fonction logistique $g(u)$:

$$g'(u) = \frac{e^u}{(1+e^u)^2} = \frac{e^u}{1+e^u} \frac{1}{1+e^u} = g(u)[1 - g(u)].$$

Ainsi :

$$\frac{\partial}{\partial \beta_0} p(x_i) = p(x_i)[1 - p(x_i)]$$
$$\frac{\partial}{\partial \beta_1} p(x_i) = x_i \, p(x_i)[1 - p(x_i)].$$

La dérivée du i-ème terme de la log-vraisemblance par rapport à β_0 est donc :

$$y_i \frac{p(x_i)[1 - p(x_i)]}{p(x_i)} - (1 - y_i) \frac{p(x_i)[1 - p(x_i)]}{1 - p(x_i)} = y_i - p(x_i)$$

et, de même, celle par rapport à β_1 est $x_i[y_i - p(x_i)]$. D'où les équations de vraisemblance :

$$\begin{cases} \dfrac{\partial}{\partial\beta_0} \ln L(\boldsymbol{\beta}) = \displaystyle\sum_{i=1}^{n} \{y_i - p(x_i)\} = 0 \\ \dfrac{\partial}{\partial\beta_1} \ln L(\boldsymbol{\beta}) = \displaystyle\sum_{i=1}^{n} \{x_i[y_i - p(x_i)]\} = 0 \end{cases}$$

ou, faisant apparaître β_0 et β_1 :

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \sum_{i=1}^{n} y_i \\ \displaystyle\sum_{i=1}^{n} x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \sum_{i=1}^{n} x_i y_i \end{cases}.$$

Ces équations n'ont pas de solution explicite et paraissent complexes. Toutefois elles ne posent pas de difficultés pour être résolues de façon itérative pour donner l'estimateur du MV $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1)$. On en déduit l'estimateur de la fonction de régression en x quelconque :

$$\widehat{p}(x) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x}}.$$

11.3.3 Matrice des variances-covariances de $\widehat{\boldsymbol{\beta}}$

En approximation on utilise les propriétés asymptotiques de l'estimateur du MV. Soit $\mathbb{I}(\boldsymbol{\beta})$ la 2×2-matrice d'information de Fisher de $\boldsymbol{\beta}$, on a alors (voir section 6.7.4) $\mathbb{V}(\widehat{\boldsymbol{\beta}}) \simeq [\mathbb{I}(\boldsymbol{\beta})]^{-1}$. Explicitons la matrice $\mathbb{I}(\boldsymbol{\beta})$. En posant $f(\mathbf{y};\boldsymbol{\beta})$ pour la fonction de probabilité conjointe du vecteur aléatoire $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ au point $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, on a (voir section 6.6.4) :

$$\mathbb{I}(\boldsymbol{\beta}) = \begin{pmatrix} -E\left(\dfrac{\partial^2}{\partial^2\beta_0} \ln f(\mathbf{Y};\boldsymbol{\beta})\right) & -E\left(\dfrac{\partial^2}{\partial\beta_0\partial\beta_1} \ln f(\mathbf{Y};\boldsymbol{\beta})\right) \\ -E\left(\dfrac{\partial^2}{\partial\beta_0\partial\beta_1} \ln f(\mathbf{Y};\boldsymbol{\beta})\right) & -E\left(\dfrac{\partial^2}{\partial^2\beta_1} \ln f(\mathbf{Y};\boldsymbol{\beta})\right) \end{pmatrix}.$$

Or $\ln f(\mathbf{y};\boldsymbol{\beta})$ n'est autre que la log-vraisemblance vue ci-dessus dont il faut calculer les dérivées partielles secondes, soit :

$$\frac{\partial^2}{\partial^2\beta_0} \ln L(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_0} \sum_{i=1}^{n} [y_i - p(x_i)] = -\sum_{i=1}^{n} \frac{\partial}{\partial\beta_0} p(x_i) = -\sum_{i=1}^{n} p(x_i)[1 - p(x_i)].$$

De même :

$$\frac{\partial^2}{\partial^2\beta_1} \ln L(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_1} \sum_{i=1}^{n} x_i[y_i - p(x_i)] = -\sum_{i=1}^{n} x_i^2\, p(x_i)[1 - p(x_i)]$$

et :

$$\frac{\partial^2}{\partial\beta_0\partial\beta_1}\ln L(\boldsymbol{\beta}) = \frac{\partial}{\partial\beta_1}\sum_{i=1}^{n}[y_i - p(x_i)] = -\sum_{i=1}^{n} x_i\, p(x_i)[1-p(x_i)]\,.$$

Ces dérivées secondes ne dépendant plus des y_i elles sont inchangées quand on passe aux espérances mathématiques[6], d'où :

$$\mathbb{I}(\boldsymbol{\beta}) = \begin{pmatrix} \displaystyle\sum_{i=1}^{n} p(x_i)[1-p(x_i)] & \displaystyle\sum_{i=1}^{n} x_i\, p(x_i)[1-p(x_i)] \\ \displaystyle\sum_{i=1}^{n} x_i\, p(x_i)[1-p(x_i)] & \displaystyle\sum_{i=1}^{n} x_i^2\, p(x_i)[1-p(x_i)] \end{pmatrix}.$$

Comme $\boldsymbol{\beta}$ est inconnu il faut estimer $\mathbb{I}(\boldsymbol{\beta})$ par $\mathbb{I}(\widehat{\boldsymbol{\beta}})$, c'est-à-dire substituer $\widehat{p}(x_i)$ à $p(x_i)$ dans l'expression ci-dessus de $\mathbb{I}(\boldsymbol{\beta})$. En inversant $\mathbb{I}(\widehat{\boldsymbol{\beta}})$ on obtient une estimation de $\mathbb{V}(\widehat{\boldsymbol{\beta}})$ que nous notons :

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} s^2(\widehat{\beta}_0) & s^2(\widehat{\beta}_0, \widehat{\beta}_1) \\ s^2(\widehat{\beta}_0, \widehat{\beta}_1) & s^2(\widehat{\beta}_1) \end{pmatrix}$$

où $s^2(\widehat{\beta}_0)$ est une estimation de la variance de $\widehat{\beta}_0$, $s^2(\widehat{\beta}_1)$ est une estimation de la variance de $\widehat{\beta}_1$ et $s^2(\widehat{\beta}_0, \widehat{\beta}_1)$ est une estimation de la covariance entre $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

Grâce à ces estimations on peut obtenir des intervalles de confiance et effectuer le test essentiel $H_0 : \beta_1 = 0$ pour décider de la significativité de la variable explicative.

11.3.4 Test $H_0 : \beta_1 = 0$

En raison de la normalité asymptotique du maximum de vraisemblance, sous H_0 la statistique $\widehat{\beta}_1 / s(\widehat{\beta}_1)$ suit approximativement une loi $\mathcal{N}(0\,;1)$ et on rejettera l'hypothèse de nullité au niveau 0,05 si sa réalisation n'est pas comprise dans l'intervalle $\pm 1{,}96$. Ce test est le *test de Wald* donné dans les logiciels. Parfois ce test est présenté avec le carré de la statistique ci-dessus dont la valeur critique doit alors être lue sur une loi $\chi^2(1)$.

On peut également envisager le test du rapport de vraisemblance généralisé fondé sur la déviance :

$$-2\left[\ln L(\widehat{\boldsymbol{\beta}}_{H_0}) - \ln L(\widehat{\boldsymbol{\beta}})\right]$$

[6] Si, comme généralement dans les modèles complexes, les dérivées secondes avaient été fonction des y_i on aurait été contraint d'estimer les espérances en prenant les expressions des dérivées secondes telles qu'elles apparaîtraient ci-dessus.

où $\widehat{\boldsymbol{\beta}}_{H_0}$ est la valeur de $\boldsymbol{\beta}$ maximisant la log-vraisemblance $\ln L(\boldsymbol{\beta})$ sous l'hypothèse H_0, c'est-à-dire avec :

$$p(x_i) = \frac{\exp \beta_0}{1 + \exp \beta_0} = p_0.$$

Seule subsiste alors la première équation de vraisemblance $\sum_{i=1}^{n}(y_i - p_0) = 0$ dont la solution est $\widehat{p}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$, la proportion de succès observée. Cette solution est naturelle puisque sous H_0 les Y_i sont des variables aléatoires de même moyenne et donc i.i.d.. On en déduit :

$$\widehat{\boldsymbol{\beta}}_{H_0} = \ln \frac{\widehat{p}_0}{1 - \widehat{p}_0}$$

qui permet de calculer la déviance ci-dessus. Celle-ci suit approximativement une loi $\chi^2(1)$ car H_0 ne spécifie qu'un seul paramètre. Ce test donne des décisions généralement en accord avec celles du test de Wald. Notons qu'il existe un autre test, appelé test du score, qui est encore plus proche du test du RV.

11.3.5 Intervalles de confiance

Toujours en vertu de la normalité asymptotique de $\widehat{\boldsymbol{\beta}}$ on peut utiliser le calcul de la matrice $\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}})$ pour établir un IC sur chaque composante de $\boldsymbol{\beta}$. Par exemple, pour β_0, on a :

$$IC_{0,95}(\beta_0) \simeq [\widehat{\beta}_0 - 1,96\ s(\widehat{\beta}_0)\,;\ \widehat{\beta}_0 + 1,96\ s(\widehat{\beta}_0)].$$

Toutefois l'intervalle de confiance qui nous intéresse le plus concerne la proportion de succès $p(x)$ pour une valeur donnée x de la valeur explicative. Considérons tout d'abord un IC sur $\beta_0 + \beta_1 x$. Asymptotiquement, son estimateur du MV $\widehat{\beta}_0 + \widehat{\beta}_1 x$ est gaussien et de variance :

$$V(\widehat{\beta}_0 + \widehat{\beta}_1 x) = V(\widehat{\beta}_0) + 2x\, cov(\widehat{\beta}_0, \widehat{\beta}_1) + x^2\, V(\widehat{\beta}_1)$$

que l'on estime par :

$$s^2(\widehat{\beta}_0 + \widehat{\beta}_1 x) = s^2(\widehat{\beta}_0) + 2x\, s^2(\widehat{\beta}_0, \widehat{\beta}_1) + x^2\, s^2(\widehat{\beta}_1).$$

D'où l'IC approché :

$$IC_{0,95}(\beta_0 + \beta_1 x) \simeq [\widehat{\beta}_0 + \widehat{\beta}_1 x - 1{,}96\, s(\widehat{\beta}_0 + \widehat{\beta}_1 x); \widehat{\beta}_0 + \widehat{\beta}_1 x + 1{,}96\, s(\widehat{\beta}_0 + \widehat{\beta}_1 x)]\,.$$

De cet intervalle on déduit celui sur $p(x)$ en appliquant aux deux bornes la fonction croissante $g(u) = \dfrac{e^u}{1 + e^u}$.

Cette procédure est la procédure duale (voir section 9.8) du test de Wald.

Exemple 11.2 Lors d'une enquête de santé publique 307 individus d'âges variant entre 18 et 85 ans ont été étudiés[7]. Parmi ceux-ci 133 souffraient d'une maladie chronique. Sachant que la proportion de personnes ayant une maladie chronique augmente avec l'âge, on envisage un modèle logistique pour estimer la probabilité d'un tel type d'affection en fonction de l'âge. La solution des deux équations de vraisemblance obtenue par un logiciel mathématique ou statistique est :

$$\begin{aligned}\widehat{\beta}_0 &= -2{,}284\\ \widehat{\beta}_1 &= 0{,}04468\,.\end{aligned}$$

La probabilité d'avoir une maladie chronique à l'âge x est donc estimée par :

$$\widehat{p}(x) = -2{,}284 + 0{,}04468\,x\,.$$

En calculant les $\widehat{p}(x_i)$ pour chacune des observations on déduit l'expression de la matrice $\mathbb{I}(\widehat{\boldsymbol{\beta}})$ qui est inversée pour donner :

$$\widehat{\mathbb{V}}(\widehat{\boldsymbol{\beta}}) = \begin{pmatrix} 0{,}1349 & -0{,}2639\times 10^{-2} \\ -0{,}2639\times 10^{-2} & 0{,}5814\times 10^{-4} \end{pmatrix}.$$

La réalisation de la statistique de Wald pour tester $H_0 : \beta_1 = 0$ est égale à $0{,}04468/\sqrt{0{,}5814\times 10^{-4}} = 5{,}86$ ce qui correspond à une P-valeur pratiquement nulle (2×10^{-9}). L'âge est donc un facteur explicatif très significatif pour la présence d'une maladie chronique.

Pour le test du RVG on a :

$$\ln L(\widehat{\boldsymbol{\beta}}_{H_0}) = \left(\sum_{i=1}^{n} y_i\right)\ln \widehat{p}_0 + \left(\sum_{i=1}^{n}(1-y_i)\right)\ln(1-\widehat{p}_0)$$

avec $\widehat{p}_0 = 133/307 = 0{,}4332$. Soit :

$$\ln L(\widehat{\boldsymbol{\beta}}_{H_0}) = 133\,\ln(0{,}4332)\; + 174\,\ln(0{,}5668) = -210{,}1\,.$$

De même on calcule :

$$\begin{aligned}\ln L(\widehat{\boldsymbol{\beta}}) &= \sum_{i=1}^{n}\{y_i \ln p(x_i) + (1-y_i)\ln[1-p(x_i)]\}\\ &= -190{,}4\end{aligned}$$

pour obtenir finalement la valeur prise par la déviance :

$$-2\left[\ln L(\widehat{\boldsymbol{\beta}}_{H_0}) - \ln L(\widehat{\boldsymbol{\beta}})\right] = 39{,}4$$

[7]Echantillon non représentatif extrait de l'enquête *Health Statistics 1990* auprès de 7200 personnes, effectuée par *Statistics Netherlands*.

qui donne aussi une P-valeur quasi nulle sur la loi $\chi^2(1)$. Notons que le carré de la statistique de Wald réalisée est $(5{,}86)^2 = 34{,}3$, ce qui est proche de la valeur obtenue par le RVG.

Donnons un IC à 95% pour β_1. On a :

$$\begin{aligned} IC_{0,95}(\beta_1) &= [0{,}04468 - 1{,}96\sqrt{0{,}5814 \times 10^{-4}}\,;\, 0{,}04468 + 1{,}96\sqrt{0{,}5814 \times 10^{-4}}] \\ &= [0{,}02974\,;\, 0{,}05963]. \end{aligned}$$

Certains logiciels statistiques donnent un IC pour e^{β_1}, soit ici $[1{,}030\,;1{,}060]$. Cette valeur, estimée ici par $e^{\widehat{\beta}_1} = 1{,}046$, a une signification particulière. Comme β_1 correspond à l'accroissement du logit de $p(x)$ quand x s'accroît d'une unité, e^{β_1} est le *rapport des chances* (odds ratio) d'une année à l'autre.

Voyons maintenant un IC sur la probabilité d'avoir une maladie chronique à l'âge de 50 ans. On a $\widehat{\beta}_0 + \widehat{\beta}_1 \times 50 = -0{,}050$ et $\widehat{p}(50) = 0{,}488$. Puis :

$$\begin{aligned} s^2(\widehat{\beta}_0 + \widehat{\beta}_1 \times 50) &= 0{,}1349 + 2 \times 50 \times (-0{,}2639)^2 + (50)^2 \times 0{,}5814 \times 10^{-4} \\ &= 0{,}01635\,. \end{aligned}$$

L'intervalle de confiance pour $\widehat{\beta}_0 + \widehat{\beta}_1 \times 50$ est donc $-0{,}050 \pm 1{,}96\sqrt{0{,}01635}$ soit $[-0{,}301\,;\, 0{,}201]$, d'où finalement :

$$IC_{0,95}(p(50)) = [0{,}425\,;0{,}550].$$

Il est intéressant de constater qu'un modèle linéaire ajusté sur la variable réponse 1/0 donne une estimation :

$$\widetilde{p}(x) = -0{,}0166 + 0{,}0101\,x$$

qui se différencie très peu de celle du modèle logistique. En effet pour $x = 50$ on obtient une probabilité identique 0,488 et pour les âges extrêmes de 20 et 80 ans on obtient respectivement 0,185 et 0,791 contre 0,199 et 0,784 pour l'approche logistique. Ceci s'explique par le fait que la plage d'âges observée se situe dans la partie centrale et quasi linéaire de la courbe logistique (voir figure 11.2). ■

11.3.6 Remarques diverses

1. La régression logistique illustre l'intérêt de la méthode du maximum de vraisemblance. Dans un modèle complexe il est peu probable de pouvoir dégager des estimateurs optimaux des paramètres. Cette méthode garantit (moyennant des conditions faibles de régularité) des estimateurs à faibles biais - et ceci d'autant plus que la taille d'échantillon est élevée - dont on peut par ailleurs estimer les variances et covariances pour construire des tests et intervalles de confiance approchés.

2. Le modèle logistique n'est évidemment pas la panacée. Une condition nécessaire, mais non suffisante, pour qu'il s'applique est que la probabilité soit de toute évidence une fonction monotone de la variable explicative. Parmi d'autres possibilités signalons les modèles *probit* et *gompit*. Le premier utilise à la place de la fonction logit $g(u)$ la fonction de répartition $\Phi(u)$ de la loi $\mathcal{N}(0\,;1)$. Celle-ci restant toutefois très proche de $g(u)$, le modèle correspondant ne se différencie pratiquement pas du modèle logistique. Le modèle gompit utilise la fonction $h(u) = 1 - \exp(-\exp(u))$ qui permet d'attribuer des probabilités plus fortes sur les extrêmes mais n'est pas symétrique. A l'instar de l'exemple précédent, bien des situations peuvent simplement être modélisées par une fonction linéaire. Même pour une réponse binaire les tests et IC vus en section 11.2 fournissent des résultats approchés corrects. Ceci confirme la forte robustesse du modèle linéaire vis-à-vis de l'hypothèse gaussienne.

3. Divers diagnostics et tests ont été proposés pour vérifier l'adéquation du modèle (notamment le test de Hosmer et Lemeshow, 2000). Une façon simple et rapide de vérifier si $p(x)$ semble bien suivre l'allure voulue est de grouper les données par classes (par exemple former des classes d'âge dans l'exemple ci-dessus), de calculer dans chaque classe la proportion de succès observée et de tracer de manière lisse la courbe passant par les points dont les abscisses sont les milieux des classes et les ordonnées sont les proportions correspondantes.

4. Comme pour la régression linéaire la régression logistique peut s'étendre à une régression multiple (plusieurs variables explicatives). Les principes de calculs sont une extension naturelle de ceux vus ci-dessus qui ne pose pas de difficultés spécifiques.

5. Si la variable explicative est catégorielle on peut appliquer la régression logistique en introduisant les variables indicatrices de chaque catégorie, sauf pour l'une d'entre elles qui sert alors de catégorie de référence. Ceci est à rapprocher de l'analyse de variance à un facteur du modèle linéaire. Toutefois, comme dans le cas linéaire, on a affaire à des interprétations particulières qui relèvent d'une méthodologie propre. A cette fin, et comme il a été indiqué en note de bas de page dans la section 11.3.1, on préfère parler de modèle logit lorsque la ou les variables explicatives sont toutes catégorielles.

Pour approfondir le sujet de la régression avec réponse binaire on pourra consulter les ouvrages : Droesbeke, Lejeune et Saporta (2004) ou, en anglais, Agresti (2002) et Chap (1998).

11.4 La régression non paramétrique

11.4.1 Introduction

Nous nous situons ici, comme en section 8.5, dans le cadre de l'estimation fonctionnelle : la fonction de régression $g(x)$ est totalement inconnue et est l'objet même à estimer. Une telle approche peut se révéler utile si l'on n'a pas d'idée précise sur une forme fonctionnelle adéquate ou si la forme de la fonction est complexe et se prête mal à une modélisation par une forme paramétrique simple.

Un avantage de la régression non paramétrique est de fournir une procédure automatique d'ajustement quel que soit le type de données. Elle est à classer parmi les méthodes dites *adaptatives.* On peut voir comme un inconvénient le fait qu'elle ne livre pas un modèle sous forme de formule facilement réutilisable pour la prévision, mais donne uniquement une description point par point de la fonction. Toutefois on pourra concevoir la procédure comme une première étape, sans aucune restriction, pour orienter ensuite la recherche d'une forme paramétrique adaptée.

Nous présenterons la méthode des noyaux qui, plus que toute autre, est une approche très intuitive du problème et qui sera en cohérence avec l'estimation de densité ou de fonction de répartition exposée en section 8.5. Par ailleurs la régression polynomiale locale, plus performante, que l'on étudiera en fin de section, en est un prolongement naturel.

Les méthodes usuelles ne concernent que les phénomènes où **la variable réponse varie en moyenne de façon lisse en fonction de la variable explicative**. De fait on fera l'hypothèse que la fonction de régression $g(x)$ est dérivable au moins à l'ordre 2. Comme pour les modèles paramétriques de ce chapitre on supposera que la variance conditionnelle de la variable Y sachant $X = x$ est indépendante de x et égale à σ^2.

11.4.2 Définition des estimateurs à noyaux

Les estimateurs à noyaux de régression (en anglais : *kernel estimators*) ont été introduits simultanément par Nadaraya (1964) et Watson (1964) qui se sont inspirés des développements accomplis dans le domaine de l'estimation de densité. Ils reposent sur une idée très intuitive proche de celle, plus ancienne, de moyenne mobile. Pour estimer $g(x)$, où x est un niveau donné de la variable explicative, à partir des réalisations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, on prend la moyenne des valeurs des y_i pour l'ensemble des observations dont les niveaux

sont situés dans un voisinage (ou fenêtre) $[x-h, x+h]$ autour de x, soit[8] :

$$\widehat{g}_n(x) = \frac{\sum_{i=1}^{n} y_i \, I_{[x-h,x+h]}(x_i)}{\sum_{i=1}^{n} I_{[x-h,x+h]}(x_i)}$$

où $I_{[x-h,x+h]}(x_i)$ est égal à 1 si $x_i \in [x-h, x+h]$ et 0 sinon. En introduisant le noyau de Rosenblatt (voir section 8.5.2) : $K(u) = 1/2$ si $-1 \leq u \leq 1$ et 0 sinon, on peut écrire :

$$\widehat{g}_n(x) = \frac{\sum_{i=1}^{n} y_i \, K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})} .$$

Cette forme se prête à la généralisation à un noyau quelconque K introduisant une moyenne pondérée des y_i. Rappelons qu'une fonction K est un noyau si elle est paire et si son intégration sur $\mathbb{R}$ donne 1. Dans le cas de la densité qui reste positive ou nulle on a imposé également que K ne soit pas négative. Cette condition est moins cruciale dans le cas de la régression (les moyennes mobiles utilisées pour le lissage des séries chronologiques comprennent d'ailleurs des coefficients négatifs afin de réduire le biais). Les propriétés de continuité et dérivabilité se transférant à la fonction estimée on aura avantage, comme pour la densité, à choisir un noyau de type biweight, par exemple, qui soit dérivable aux bornes du support.

11.4.3 Biais et variance

Comme en régression linéaire nous distinguerons deux cas, selon que les x_i sont déterminés par un plan d'expérience ou qu'ils sont réalisations de variables aléatoires X_i.

Cas des x_i fixés

L'estimateur $\widehat{g}_n(x)$ en un point x donné étant une fonction linéaire des Y_i, le biais et la variance se calculent aisément. Posant $w_i = K(\frac{x-x_i}{h})$ on a :

[8] La plupart des propriétés que nous expliciterons sont de nature asymptotique. C'est pourquoi nous indiçons l'estimation $\widehat{g}_n(x)$ par n comme pour la densité. De plus nous utiliserons la même notation pour estimation et estimateur.

$$E(\widehat{g}_n(x)) - g(x) = \frac{\sum_{i=1}^{n} w_i(g(x_i) - g(x))}{\sum_{i=1}^{n} w_i}$$

$$V(\widehat{g}_n(x)) = \sigma^2 \frac{\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2} .$$

Pour fixer les idées prenons le cas de valeurs x_i espacées régulièrement sur l'intervalle d'intérêt $[a, b]$. Si n est assez grand on peut approcher la somme finie par une intégrale pour obtenir (voir Gasser et Müller, 1984) :

$$E(\widehat{g}_n(x)) - g(x) = \int_{-1}^{1} K(u)[g(x + uh) - g(x)]du + o(\frac{1}{nh})$$
$$V(\widehat{g}_n(x)) = \frac{b-a}{nh}\sigma^2 \int_{-1}^{1} [K(u)]^2 du + O(\frac{1}{n^2}) .$$

Ces formules sont très semblables à celles de la densité et conduisent aux mêmes conditions nécessaires pour la convergence en moyenne quadratique, à savoir $n \to \infty$, $h \to 0$, $nh \to \infty$. Pour préciser le comportement du biais prenons le développement de Taylor :

$$g(x + uh) = g(x) + uhg'(x) + \frac{1}{2}(uh)^2 g''(x) + O(h^3)$$

d'où :

$$E(\widehat{g}_n(x)) - g(x) = \frac{h^2}{2} g''(x) \int_{-1}^{1} u^2 K(u)du + O(h^3) + o(\frac{1}{nh}).$$

A partir de ces formules asymptotiques il est possible d'étudier la vitesse de convergence de l'e.q.m. pour la valeur de h optimale en x fixé. Les développements sont analogues à ceux de la densité et l'on trouve également une vitesse optimale de l'ordre de $n^{-4/5}$ avec h de l'ordre de $n^{-1/5}$.

Toutefois ces formules ne sont valables que si la fenêtre $[x - h, x + h]$ est intégralement contenue dans l'intervalle $[a, b]$. Si, par exemple, $x - h < a$ avec

$x - qh = a$ où $0 \leq q < 1$, le terme $uhg'(x)$ du développement de Taylor fournit un premier terme de biais $hg'(x) \int_{-q}^{1} K(u)du$ généralement non nul. Supposons que la fonction de régression g soit pratiquement linéaire au voisinage de x. Si la fenêtre est à l'intérieur de $[a, b]$ le biais est nul par compensation de part et d'autre de x, mais si $x - h < a$ la partie gauche de la fenêtre contient moins de points ce qui introduit le biais.

Notons aussi que si g admet un extremum en x alors le biais est du signe de $g''(x)$ ce qui entraîne un phénomène d'écrêtement déjà rencontré pour l'estimation de densité. Ici aussi l'on peut introduire des noyaux d'ordre 4 contenant nécessairement une plage négative (voir les remarques diverses en fin de section 8.5.2) pour remédier à ce problème. C'est d'ailleurs pour cette même raison que les moyennes mobiles utilisent des poids négatifs sur les extrémités.

Cas des X_i aléatoires

On supposera que la densité conjointe $f_{X,Y}$ du couple (X, Y) est continue dans $\mathbb{R}^2$. On peut alors montrer que $\widehat{g}_n(x)$ converge en probabilité vers $g(x)$ en tout point x tel que $f_X(x) \neq 0$. Collomb (1977) a établi les formules asymptotiques du biais et de la variance suivantes (sous certaines conditions de régularité de $f_{X,Y}$ et de K) :

$$E(\widehat{g}_n(x)) - g(x) = h^2 \int_{-1}^{1} u^2 K(u)du \left[g'(x) \frac{f'_X(x)}{f_X(x)} + \frac{1}{2} g''(x) \right] + o(h^2)$$

$$V(\widehat{g}_n(x)) = \frac{\sigma^2}{nh} \frac{1}{f_X(x)} \int_{-1}^{1} [K(u)]^2 du + o(\frac{1}{nh}) \,.$$

Ces formules d'approximation asymptotique restent très théoriques car en pratique les valeurs de h convenables sont loin d'être faibles même avec des grands échantillons. De plus, elles n'intègrent pas les effets de bord. Néanmoins elles reflètent bien les écueils des estimateurs à noyaux. Dans le cas où X suit une loi continue uniforme sur $[a, b]$ on retrouve les formules et les problèmes précédents. Si la loi de X n'est pas uniforme il s'introduit un terme de biais supplémentaire $h^2 \int_{-1}^{1} u^2 K(u)du \, .g'(x) \frac{f'_X(x)}{f_X(x)}$, même si g est linéaire, dû au gradient de densité autour de x mis en évidence par $f'_X(x)$, lequel déséquilibre la symétrie.

La régression par noyau présente, nous venons de le voir, des inconvénients majeurs : effets de bord (importants vu les largeurs de fenêtre nécessaires à un lissage satisfaisant), écrêtement des extrema, présence de biais même pour une fonction de régression linéaire si la densité de X n'est pas uniforme ou, dans le cas d'un plan d'expérience, si la répartition des x_i n'est pas régulière. La méthode qui suit va y remédier.

11.4.4 La régression polynomiale locale

Cette méthode (RPL) est une généralisation de la méthode *Local Weighted Regression* ou LOWESS de Cleveland (1979) proposée par Lejeune[9] (1983) dans le cadre de l'estimation par noyau. Elle consiste, pour estimer $g(x)$, à ajuster une fonction polynomiale de degré s choisi, sur les couples de points (x_i, y_i) dont les x_i sont situés dans le voisinage (fenêtre) $[x-h, x+h]$ de x. L'ajustement s'entend au sens classique des moindres carrés des résidus $y_i - \widehat{y}_i$ (voir section 11.2) et est donc un cas particulier de régression linéaire multiple, la fonction polynomiale étant linéaire **par rapport aux paramètres inconnus.** Il se résout matriciellement comme indiqué en section 11.2.7. Alors $g(x)$ est estimé par la valeur ajustée au point x que nous noterons $\widetilde{g}_n(x)$.

Soit $P(u) = a_0 + a_1 u + \cdots + a_s u^s$ un polynôme de degré s. Cet ajustement s'opère avec les valeurs de $a_0, a_1, \ldots, a_s$ telles que l'expression :

$$\sum_{i=1}^{n} \left[y_i - (a_0 + a_1 x_i + \cdots + a_s x_i^s)\right]^2 K\left(\frac{x - x_i}{h}\right)$$

soit minimale, où K est la fonction indicatrice de l'appartenance de la valeur x_i à la fenêtre ($K(u) = 1$ si $|u| \leq 1$, 0 sinon). Soit $\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_s$ les valeurs permettant d'atteindre le minimum, $g(x)$ est alors estimé par $\widetilde{g}_n(x) = \widehat{a}_0 + \widehat{a}_1 x + \cdots + \widehat{a}_s x^s$.

Comme pour l'estimateur à noyau les propriétés de la fonction K se transfèrent à la fonction $\widetilde{g}_n$ et l'on aura avantage à substituer à la fonction indicatrice une fonction de pondération dérivable partout, ce qui conduit à une solution des *moindres carrés pondérés.* On montre aisément que la solution matricielle de la section 11.2.7 pour le vecteur des paramètres devient $(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{Y}$ où $\mathbf{W}$ est la $n \times n-$matrice diagonale des poids $K(\frac{x-x_i}{h})$ affectés aux n observations. On pourra choisir pour K le noyau le plus simple possédant les qualités requises, à savoir le biweight de Tukey. Notons d'ailleurs que l'estimation par noyau correspond au cas particulier $s = 0$, car la valeur de a_0 qui minimise $\sum_{i=1}^{n} [y_i - a_0]^2 K(\frac{x-x_i}{h})$ est la moyenne pondérée des y_i avec les poids $K(\frac{x-x_i}{h})$.

Intuitivement on peut voir les avantages de la méthode pour autant que s soit supérieur à 0. En effet on perçoit bien qu'avec un simple ajustement local linéaire ($s = 1$) on doit pouvoir prendre en compte les problèmes résultant du différentiel de densité de points de part et d'autre de la valeur x, en particulier le problème des effets de bord. De même, mais avec $s = 2$, il est possible d'obtenir un meilleur ajustement pour les zones à forte courbure, notamment en ce qui concerne le problème de l'écrêtement des extrema. Nous allons vérifier cela sur les propriétés de la RPL.

[9] A l'origine la méthode a été également appelée «régression polynomiale mobile» par référence à la moyenne mobile analogue dans son esprit à l'estimateur à noyau.

Cas des x_i fixés

On démontre (voir Lejeune, 1984, 1985) que la RPL au degré s :

1. produit, en tout x fixé, un biais en $O(h^{s+1})$ quelle que soit la répartition des x_i et quelle que soit la fonction de poids utilisée,
2. est, avec des pondérations uniformes et pour une largeur de fenêtre fixée, l'estimateur de variance minimale de $g(x)$ parmi les estimateurs linéaires (en fonction des Y_i) dont le biais est en $O(h^{s+1})$.

La première propriété montre le caractère adaptatif de la RPL pour ce qui concerne le problème du biais. Quant à l'optimalité exprimée dans la seconde propriété elle doit être quelque peu sacrifiée si l'on veut bénéficier de la dérivabilité de la fonction $\widetilde{g}_n$. Toutefois l'incidence est faible car la fonction de poids n'influe pas de façon très sensible sur la variance.

Cas des X_i aléatoires

Le biais et la variance asymptotiques ont été établis par Fan (1993). En fait la variance ne dépend pas du degré de la RPL et reste égale à celle indiquée plus haut pour l'estimateur à noyau correspondant au cas $s = 0$. Pour le biais on obtient pour $s \geq 1$:

$$E(\widetilde{g}_n(x)) - g(x) = \frac{h^{s+1}}{(s+1)!} g^{(s+1)}(x) \int_{-1}^{1} u^{s+1} K(u) du \; + o(h^{s+1}).$$

Par rapport à l'estimateur à noyau classique on constate qu'avec $s = 1$ le terme dû au gradient de densité autour de x (mis en évidence par le facteur $g'(x) f'_X(x)/f_X(x)$) disparaît.

En pratique le choix $s = 1$, proposé à l'origine par Cleveland et repris par divers auteurs, n'est cependant pas satisfaisant car il ne traite pas le problème de l'écrêtement des extrema (ou, plus généralement, du biais dans les zones à forte courbure). Pour cela on peut considérer qu'un ajustement parabolique ($s = 2$) suffira, d'autant qu'à n fini la variance augmente avec l'ordre du biais. La figure 11.3 illustre la bonne qualité d'un tel ajustement. Notons que la RPL n'évite pas le problème du choix de la largeur de fenêtre. Néanmoins on constate que l'estimation de $g(x)$ est moins sensible à ce paramètre qu'avec un estimateur à noyau.

Pour approfondir la régression non paramétrique on pourra consulter les ouvrages de Haerdle (1990) et de Simonoff (1996).

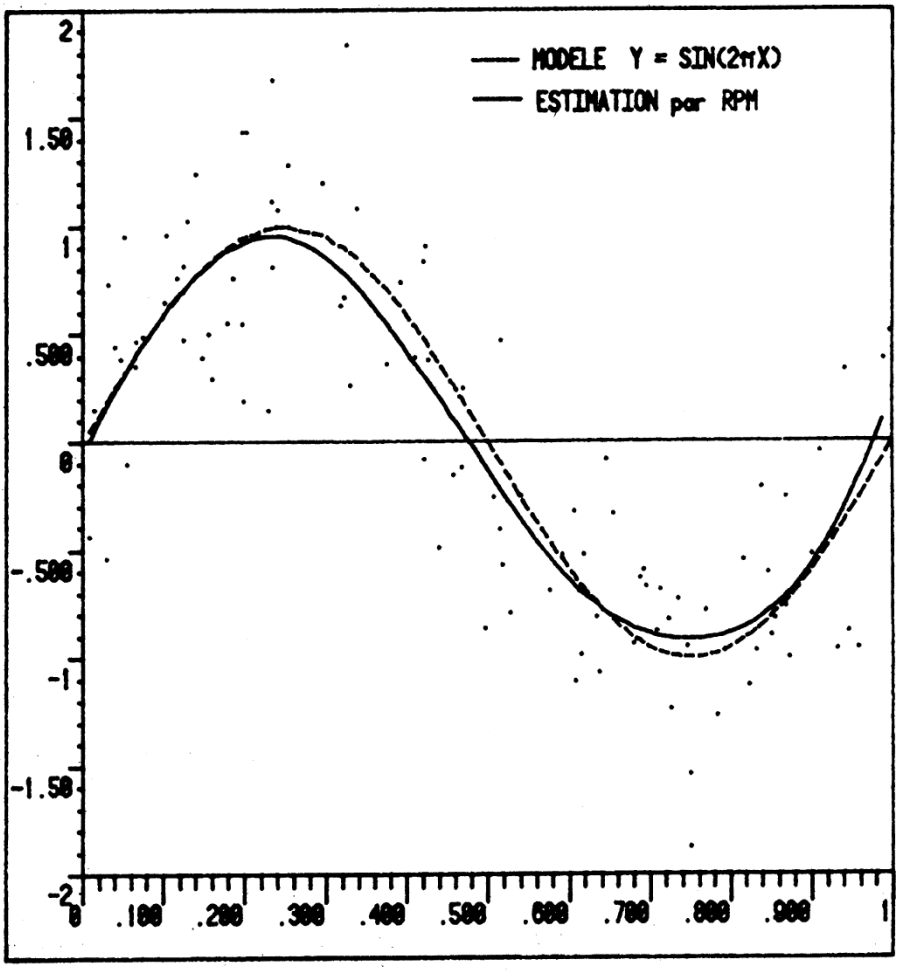


Figure 11.3 - Estimation par régression polynomiale locale de degré 2 avec pondérations biweight et $h = 0,4$: échantillon de 100 observations simulées par un modèle sinusoïdal à erreurs $\mathcal{N}(0\,;4)$ avec abcisses $\mathcal{U}[0,1]$. *Reproduction autorisée de la Revue de Statistique Appliquée, vol. XXXIII, n°3*, page 62, 1985.

11.5 Exercices

Exercice 11.1 Soit (X, Y) un couple aléatoire gaussien de paramètres μ_X, μ_Y, σ_X^2, σ_Y^2 et ρ.
Partant de l'expression matricielle générale de la section 3.9 développer l'expression analytique de la densité conjointe du couple.

Montrer que la loi conditionnelle de Y sachant $X = x$ est gaussienne de moyenne $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ et de variance $\sigma_Y^2(1 - \rho^2)$.

Aide : on utilisera le résultat de la section 3.2 pour la densité conditionnelle : $f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

Exercice 11.2 * Soit (X, Y) un couple aléatoire non nécessairement gaussien de moyennes, variances et corrélation linéaire μ_X, μ_Y, σ_X^2, σ_Y^2 et ρ. Montrer que si $E(Y|X = x)$ est une fonction linéaire de x alors cette fonction est $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$.

Aide : Soit $\varphi(x) = E(Y|X = x)$. Calculer $E(\varphi(X))$ d'une part de façon générale en passant par la formule pour la densité conditionnelle :

$$E(Y|X = x) = \int_{\mathbb{R}} y\frac{f_{X,Y}(x, y)}{f_X(x)}dy$$

et d'autre part par l'expression linéaire $E(Y|X = x) = \beta_0 + \beta_1 x$ pour obtenir une première équation en β_0 et β_1, puis calculer de même $E(X\varphi(X))$ pour obtenir une deuxième équation.

Montrer que si, de plus, la variance conditionnelle $V(Y|X = x)$ ne dépend pas de x alors elle est égale à $\sigma_Y^2(1 - \rho^2)$.

Aide : Soit $\psi(x) = V(Y|X = x)$. Calculer $E(\psi(X))$ d'une part en décentrant $V(Y|X = x)$ et d'autre part en tenant compte de la propriété sur $V(Y|X = x)$.

Exercice 11.3 Démontrer la formule de décomposition de la somme des carrés totale pour la régression linéaire simple.

Aide : partir de $y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})$ et montrer, en remplaçant $\widehat{y}_i$ par $\overline{y} + \widehat{\beta}_1(x_i - \overline{x})$, que $\sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = 0$.

Exercice 11.4 *(intervalle de prédiction)* Dans le cadre de la régression linéaire gaussienne simple on cherche à prévoir une observation Y_0 pour le niveau x_0 de la variable explicative.
Montrer que $Y_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$ suit une loi de Gauss de moyenne 0 et de variance $\sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}\right)$ (aide : Y_0 est indépendante des Y_i sur lesquels reposent $\widehat{\beta}_0$ et $\widehat{\beta}_1$).
En déduire que :

$$P\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - t_{0,975}^{(n-2)}\, S\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}} < Y_0 < \right.$$
$$\left. \widehat{\beta}_0 + \widehat{\beta}_1 x_0 + t_{0,975}^{(n-2)}\, S\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}}\right) = 0,95.$$

Les réalisations des bornes d'encadrement de Y_0 (qui sont aléatoires) constituent un «intervalle de prédiction à 95%» pour Y_0.

Exercice 11.5 * Démontrer le théorème 11.1 pour $\widehat{\beta}_1$ (cet exercice nécessite la connaissance de la méthode du multiplicateur de Lagrange).

Aide : établir les deux contraintes sur les a_i pour qu'un estimateur de la forme $\sum_{i=1}^{n} a_i Y_i$ soit sans biais pour β_1. Puis minimiser la variance d'un tel estimateur sous ces contraintes.

Exercice 11.6 Pour la régression simple, déterminer la statistique λ du test du rapport de vraisemblance généralisé pour l'hypothèse $H_0 : \beta_1 = 0$ et montrer que c'est une fonction décroissante de la statistique F.

Corrigés des exercices

Chapitre 1 : Variables aléatoires

Exercice 1.1

Clairement, on a (avec la convention $A_0 = \emptyset$) :

$$\bigcup_{j=1}^{n} A_j = A_n = \bigcup_{j=1}^{n} (A_j \cap \overline{A}_{j-1})$$

et :

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} (A_n \cap \overline{A}_{n-1}) .$$

La suite $\{A_n \cap \overline{A}_{n-1}\}$ étant une suite d'événements incompatibles, on a :

$$\begin{aligned} P(\bigcup_{n=1}^{\infty} A_n) &= \sum_{n=1}^{\infty} P(A_n \cap \overline{A}_{n-1}) \\ &= \lim_{n\to\infty} \sum_{j=1}^{n} P(A_j \cap \overline{A}_{j-1}) \\ &= \lim_{n\to\infty} P\left(\bigcup_{j=1}^{n} (A_j \cap \overline{A}_{j-1}) \right) \\ &= \lim_{n\to\infty} P(A_n) . \end{aligned}$$

Note : Soit X une v.a. de fonction de répartition F_X, montrons que $P(X < x) = F_X(x^-)$. Considérons l'événement $A_n =]-\infty, x - \frac{1}{n}]$ – soit, avec la convention de notation usuelle $(X \leq x - \frac{1}{n})$. Comme $\bigcup_{n=1}^{\infty} A_n =]-\infty, x[$, on a : $P(X < x) = \lim_{n\to\infty} P(X \leq x - \frac{1}{n}) = \lim_{n\to\infty} F_X(x - \frac{1}{n})$. Or, en raison de la non-décroissance de F_X, $\lim_{n\to\infty} F_X(x - \frac{1}{n}) = \lim_{\varepsilon\to 0} F_X(x - \varepsilon) = F_X(x^-)$.

Exercice 1.2

Utilisons le fait que $\{\overline{B}_n\}$ est une suite croissante. On a :

$$\begin{aligned} P(\bigcap_{n=1}^{\infty} B_n) &= 1 - P(\bigcup_{n=1}^{\infty} \overline{B}_n) \\ &= 1 - \lim_{n\to\infty} P(\overline{B}_n) \\ &= 1 - \lim_{n\to\infty} [1 - P(B_n)] = \lim_{n\to\infty} P(B_n). \end{aligned}$$

Note : Dans le même contexte d'application que celui de la note précédente, prenons $B_n =]-\infty, x+\frac{1}{n}]$. On a ainsi $\lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} F_X(x+\frac{1}{n}) = F_X(x^+)$. Comme $\bigcap_{n=1}^{\infty} B_n =]-\infty, x]$ il découle du résultat précédent que $F_X(x^+) = F_X(x)$, ce qui établit la continuité à droite de la fonction de répartition.

Exercice 1.3

Considérons la suite croissante d'événements $\{A_n\}$ avec $A_n =]-\infty, n]$. Comme $\bigcup_{n=1}^{\infty} A_n = \mathbb{R}$, on a :

$$P(\bigcup_{n=1}^{\infty} A_n) = 1 = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} F_X(n).$$

En raison de la non décroissance de F_X, on a $\lim_{n\to\infty} F_X(n) = \lim_{x\to\infty} F_X(x) = F_X(+\infty) = 1$.

De même avec $A_n =]-n, +\infty]$, on a $P(\bigcup_{n=1}^{\infty} A_n) = 1 = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} [1 - F_X(-n)]$, d'où $\lim_{n\to\infty} [F_X(-n)] = 0$ soit $F_X(-\infty) = 0$.

Exercice 1.4

Considérons la suite décroissante d'événements $\{B_n\}$ avec $B_n =]x-\frac{1}{n}, x]$. Notons que pour tout n, B_n est un bien événement puisque :

$$]x - \frac{1}{n}, x] =]-\infty, x] \cap \overline{]-\infty, x - \frac{1}{n}]}.$$

Or $\bigcap_{n=1}^{\infty} B_n = \{x\}$, événement dont la probabilité se note conventionnellement $P(X = x)$. D'où, d'après le résultat de l'exercice 1.2 :

$$\begin{aligned} P(X = x) &= \lim_{n\to\infty} P(B_n) = \lim_{n\to\infty} P(x - \frac{1}{n} < X \leq x) \\ &= \lim_{n\to\infty} \left[F_X(x) - F_X(x - \frac{1}{n}) \right] \\ &= F_X(x) - \lim_{n\to\infty} \left[F_X(x - \frac{1}{n}) \right] = F_X(x) - F_X(x^-). \end{aligned}$$

Note : Dans la note de l'exercice 1.1, on a établi que $P(X < x) = F_X(x^-)$. Comme $(X \leq x) = (X < x) \cup (X = x)$ le résultat ci-dessus est alors immédiat.

Exercice 1.5

La fonction de probabilité p_X de X est définie sur $\mathbb{N}^* = \{1, 2, 3, \cdots\}$.
$p_X(1) = \frac{1}{6}$, $p_X(2) = \frac{5}{6}\frac{1}{6}$, $p_X(3) = \left(\frac{5}{6}\right)^2 \frac{1}{6}$ et plus généralement :

$$p_X(k) = \left(\frac{5}{6}\right)^{k-1} \frac{1}{6} \quad \text{pour } k \in \mathbb{N}^*.$$

On a bien : $\sum_{k\in\mathbb{N}^*} p_X(k) = \frac{1}{6}\left[1 + \frac{5}{6} + \left(\frac{5}{6}\right)^2 + \cdots\right] = \frac{1}{6}\left[\frac{1}{1-\frac{5}{6}}\right] = 1$.
$P(1 < X \leq 3) = p_X(2) + p_X(3) = \frac{5}{6}\frac{1}{6} + \left(\frac{5}{6}\right)^2 \frac{1}{6} \simeq 0,255$.
Pour $k \in \mathbb{N}^*$ on a :

$$P(X > k) = \left(\frac{5}{6}\right)^k \frac{1}{6}\left[1 + \frac{5}{6} + \left(\frac{5}{6}\right)^2 + \cdots\right] = \left(\frac{5}{6}\right)^k.$$

Donc $F_X(k) = 1 - \left(\frac{5}{6}\right)^k$ et plus généralement :

$$F_X(x) = \begin{cases} 0 & \text{pour } x < 1 \\ 1 - \left(\frac{5}{6}\right)^k & \text{pour } x \in [k, k+1[\end{cases}$$

Son graphe est une fonction en escalier avec sauts des marches aux valeurs $1, 2, 3, \cdots$. Il y a continuité à droite (voir figure 1.1).

Note : Il s'agit d'une variante de la loi géométrique (section 4.1.4)

Exercice 1.6

On doit avoir d'abord $f(x) \geq 0$ pour $x \in [0, 1]$ soit $c > 0$ puisque $x(1-x) \geq 0$ sur $[0, 1]$. Puis :

$$\int_0^1 cx(1-x)dx = c\int_0^1 (x - x^{2)}dx = c\left[\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1 = c\frac{1}{6}$$

doit être égal à 1. D'où $c = 6$. La fonction de répartition vaut, pour $x \in [0, 1]$: $6\int_0^x (x - x^{2)}dx = 6\left[\frac{x^2}{2} - \frac{x^3}{3}\right]$. Plus généralement :

$$F(x) = \begin{cases} 0 & \text{pour } x \leq 0 \\ 6\left[\frac{x^2}{2} - \frac{x^3}{3}\right] & \text{pour } x \in [0, 1] \\ 1 & \text{pour } x \geq 1 \end{cases}$$

On vérifie que F est continue partout. La médiane est la valeur de x telle que $F(x) = \frac{1}{2}$. L'équation n'est pas simple à résoudre mais l'on peut constater que le graphe de $f(x)$ est symétrique par rapport à $x = \frac{1}{2}$ qui est donc la médiane (et la moyenne).

Note : Il s'agit de la loi bêta $Beta(1, 1)$ (section 4.2.9).

Exercice 1.7

La fonction $F(x)$ est non décroissante, elle part de 0 et tend vers 1 quand x tend vers $+\infty$. Comme elle est continue partout, elle caractérise une v.a. continue. Elle est strictement croissante sur le support $[0, +\infty[$ de la loi et les quantiles sont donc uniques. Le premier quartile $x_{0,25}$ vérifie $1-e^{-\frac{x_{0,25}}{2}} = 0,25$, d'où $x_{0,25} = -2\ln(1-0,25) \simeq 0,575$. De même $x_{0,75} = -2\ln(1-0,75) \simeq 2,77$.

$P(1 < X \leq 2) = F(2) - F(1) = e^{-\frac{1}{2}} - e^{-\frac{2}{2}} \simeq 0,239$.

Note : Il s'agit de la loi exponentielle $\mathcal{E}(\frac{1}{2})$ (section 4.2.2).

Exercice 1.8

Posons $Y = 1/X$. Au support $[0, 1]$ de X correspond le support $[1, +\infty[$ de Y. Alors, pour $y \in [1, +\infty[$, on a :

$$F_Y(y) = P(Y \leq y) = P(\frac{1}{X} \leq y) = P(X \geq \frac{1}{y}) = 1 - F_X(\frac{1}{y}).$$

Or, pour $t \in [0, 1]$, $F_X(t) = \int_0^t 2x dx = t^2$. D'où :

$$F_Y(y) = \begin{cases} 0 & \text{pour } y \leq 1 \\ 1 - \frac{1}{y^2} & \text{pour } y \geq 1 \end{cases}.$$

On vérifie la continuité de F_Y au point $y = 1$.

Posons $Z = \ln(1/X)$. Au support $[0, 1]$ de X correspond le support $[0, +\infty[$ de Z. Alors, pour $z \in [0, +\infty[$, on a :

$$F_Z(z) = P(Z \leq z) = P(\ln(\frac{1}{X}) \leq z) = P(X \geq \frac{1}{e^z}) = 1 - F_X(\frac{1}{e^z}).$$

D'où :

$$F_Z(z) = \begin{cases} 0 & \text{pour } z \leq 0 \\ 1 - e^{-2z} & \text{pour } z \geq 0 \end{cases}.$$

On vérifie la continuité de F_Z au point $z = 0$. Il s'agit de la loi exponentielle $\mathcal{E}(2)$ (section 4.2.2).

Exercice 1.9

Sur $[0, 1]$ on a $F_X(x) = x$. Au support $[0, 1]$ de X correspond le support $[0, +\infty[$ de Y. Alors, pour $y \in [0, +\infty[$, on a :
$F_Y(y) = P(Y \leq y) = P(-\theta \ln(1 - X) \leq y) = P(\ln(1 - X) \geq -\frac{y}{\theta})$
$F_Y(y) = P(X \leq 1 - e^{-\frac{y}{\theta}}) = 1 - e^{-\frac{y}{\theta}}$.
Pour $y \leq 0$ $F_Y(y)$ vaut 0. On vérifie la continuité de F_Y au point $y = 0$.

Note : Il s'agit de la loi exponentielle $\mathcal{E}(\frac{1}{\theta})$ (section 4.2.2). La transformation de nombres au hasard par la fonction $-\theta \ln(1 - x)$ permet donc de simuler des observations d'une loi exponentielle $\mathcal{E}(\frac{1}{\theta})$ (voir section 4.3).

Chapitre 2 : Espérance mathématique et moments

Exercice 2.1

$E(Y) = (0^2 - 1) \times 0,7 + (1^2 - 1) \times 0,2 + (2^2 - 1) \times 0,1 = -0,4.$

Exercice 2.2

Suivons la méthode de la section 1.6.

$$F_Z(z) = P(Z \leq z) = P(g(X) \leq z).$$

Comme g est croissante, l'événement $(g(X) \leq z)$ est identique à l'événement $(X \leq g^{-1}(z))$, donc :

$$F_Z(z) = P(X \leq g^{-1}(z)) = F_X(g^{-1}(z)).$$

Par dérivation, on obtient :

$$f_Z(z) = F'_Z(z) = F'_X(g^{-1}(z))(g^{-1}(z))' = f_X(g^{-1}(z))(g^{-1}(z))'$$

ce qui peut s'écrire, puisque g^{-1} est également croissante et $(g^{-1})'$ est positive : $f_Z(z) = f_X(g^{-1}(z))|(g^{-1}(z))'|$.

Si g est décroissante, l'événement $(g(X) \leq z)$ est identique à $(X \geq g^{-1}(z))$, donc :

$$F_Z(z) = P(X \geq g^{-1}(z)) = 1 - F_X(g^{-1}(z)).$$

Par dérivation, $f_Z(z) = F'_Z(z) = -F'_X(g^{-1}(z))(g^{-1}(z))' = -f_X(g^{-1}(z))(g^{-1}(z))'$. Comme g^{-1} est décroissante, $(g^{-1})'$ est négative et $f_Z(z) = f_X(g^{-1}(z))|(g^{-1}(z))'|$ également.

Soit maintenant g strictement croissante, variant de a à b quand x croît de $-\infty$ à $+\infty$ (où a et b peuvent, respectivement, être $-\infty$ ou $+\infty$). Alors :

$$E(Z) = \int_a^b z\, f_X(g^{-1}(z))(g^{-1}(z))' dz.$$

Effectuons le changement de variable $x = g^{-1}(z)$ avec $dx = (g^{-1}(z))'dz$ et $z = g(x)$, d'où immédiatement :

$$E(Z) = \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx.$$

Soit g strictement décroissante. Supposons que $g(x)$ décroît de b à a $(a < b)$ quand x croît de $-\infty$ à $+\infty$ (où a et b peuvent, respectivement, être $-\infty$ ou $+\infty$). Alors :

$$E(Z) = \int_{a}^{b} -z\, f_X(g^{-1}(z))(g^{-1}(z))'\, dz$$

et par le changement de variable $x = g^{-1}(z)$ on obtient :

$$E(Z) = \int_{+\infty}^{-\infty} -g(x)\, f_X(x)\, dx = \int_{+\infty}^{-\infty} g(x)\, f_X(x)\, dx.$$

Exercice 2.3

Soit X de densité $f(x)$, alors :

$$\begin{aligned}
\Psi(t) = E(e^{tX}) &= \int_{-\infty}^{+\infty} e^{tx} \frac{1}{2} e^{-|x|}\, dx \\
&= \int_{-\infty}^{0} e^{tx} \frac{1}{2} e^{x}\, dx + \int_{0}^{+\infty} e^{tx} \frac{1}{2} e^{-x}\, dx \\
&= \frac{1}{2} \int_{-\infty}^{0} e^{(t+1)x}\, dx + \frac{1}{2} \int_{0}^{+\infty} e^{(t-1)x}\, dx \\
&= \frac{1}{2(t+1)} \left[e^{(t+1)x} \right]_{-\infty}^{0} + \frac{1}{2(t-1)} \left[e^{(t-1)x} \right]_{0}^{+\infty} \\
&= \frac{1}{2(t+1)} \left[1 - \lim_{x \to -\infty} e^{(t+1)x} \right] + \frac{1}{2(t-1)} \left[\lim_{x \to +\infty} e^{(t-1)x} - 1 \right]
\end{aligned}$$

La première limite est 0 si $t + 1 > 1$, soit $t > -1$, et la deuxième est 0 si $t < 1$. Donc $\Psi(t)$ est définie au voisinage de 0 et :

$$\Psi(t) = \frac{1}{2(t+1)} - \frac{1}{2(t-1)} = \frac{(t-1) - (t+1)}{2(t^2-1)} = \frac{1}{1-t^2} \text{ pour } |t| < 1.$$

En développant $\Psi(t)$ en série entière (voir Note 2.3) :

$$\Psi(t) = 1 + t^2 + t^4 + \cdots ,$$

on accède directement aux moments, d'où : $\mu_2 = 2! = 2$, $\mu_4 = 4! = 24$. Ce sont aussi les moments centrés, donc le coefficient d'aplatissement vaut :

$$\frac{\mu'_4}{\sigma^4} - 3 = \frac{24}{4} - 3 = 3$$

ce qui indique un pic plus pointu en la moyenne que pour la loi de Gauss.

Exercice 2.4

On a :

$$E(X) = \int_a^{+\infty} x \, \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx = \theta a^\theta \int_a^{+\infty} \frac{1}{x^\theta} \, dx.$$

L'intégrale ci-dessus ne convergeant que si $\theta > 1$, la moyenne n'existe qu'à cette condition et vaut alors :

$$\begin{aligned} E(X) &= \theta a^\theta \left[\frac{x^{-\theta+1}}{-\theta+1}\right]_a^{+\infty} \\ &= \theta a^\theta \left[\frac{a^{-\theta+1}}{\theta-1}\right] = \frac{a\theta}{\theta-1}. \end{aligned}$$

Puis :

$$E(X^2) = \int_a^{+\infty} x^2 \, \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx = \theta a^\theta \int_a^{+\infty} \frac{1}{x^{\theta-1}} \, dx.$$

L'intégrale ci-dessus ne convergeant que si $\theta > 2$, $E(X^2)$ et la variance n'existent qu'à cette condition. Alors :

$$\begin{aligned} E(X^2) &= \theta a^\theta \left[\frac{x^{-\theta+2}}{-\theta+2}\right]_a^{+\infty} \\ &= \theta a^\theta \left[\frac{a^{-\theta+2}}{\theta-2}\right] = \frac{\theta a^2}{\theta-2}. \end{aligned}$$

Ainsi :

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 = \frac{\theta a^2}{\theta-2} - \frac{\theta^2 a^2}{(\theta-1)^2} \\ &= \frac{\theta a^2(\theta-1)^2 - \theta^2 a^2(\theta-2)}{(\theta-2)(\theta-1)^2} = \frac{\theta a^2[\theta^2 - 2\theta + 1 - \theta(\theta-2)]}{(\theta-2)(\theta-1)^2} \\ &= \frac{\theta a^2}{(\theta-2)(\theta-1)^2} \end{aligned}$$

Généralisons à $\mu_r = E(X^r)$ avec $r > 2$, r entier :

$$E(X^r) = \int_a^{+\infty} x^r \, \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx = \theta a^\theta \int_a^{+\infty} \frac{1}{x^{\theta-r+1}} \, dx.$$

L'intégrale ne converge que si $\theta - r + 1 > 1$, soit $\theta > r$. Notons qu'il en va de même pour $\mu_r^{'} = E((X-\mu)^r)$ qui s'exprime en fonction de $\mu_r, \mu_{r-1}, \cdots, \mu_2, \mu$. Or si μ_r existe, les moments d'ordres inférieurs existent. Pour $\theta > r$ on a donc :

$$\begin{aligned} E(X^r) &= \theta a^\theta \left[\frac{x^{-\theta+r}}{-\theta+r}\right]_a^{+\infty} = \theta a^\theta \left[\frac{a^{-\theta+r}}{\theta-r}\right] \\ &= \frac{\theta a^r}{\theta - r}. \end{aligned}$$

La condition $\theta > r$ laisse entendre que la fonction génératrice − permettant pour θ fixé d'obtenir tous les moments − ne peut exister, ce que nous vérifions directement en calculant :

$$\Psi(t) = E(e^{tX}) = \int_a^{+\infty} e^{tx} \, \frac{\theta}{a} \left(\frac{a}{x}\right)^{\theta+1} dx = \theta a^\theta \int_a^{+\infty} \frac{e^{tx}}{x^{\theta+1}} \, dx.$$

Or, quel que soit $\theta \in \mathbb{R}$, $\lim\limits_{x \to +\infty} \frac{e^{tx}}{x^{\theta+1}} \to +\infty$ si $t > 0$ et l'intégrale ne peut converger. Donc Ψ n'est définie dans aucun voisinage de 0, condition nécessaire pour la définition de la fonction génératrice des moments afin que $\Psi'(0), \Psi^{''}(0)$, etc., puissent exister.

Exercice 2.5

1.

$$E(\frac{1}{X}) = \int_0^1 \frac{1}{x} \, 3x^2 \, dx = 3 \int_0^1 x \, dx = 3 \left[\frac{x^2}{2}\right]_0^1 = \frac{3}{2}.$$

2. X prend ses valeurs sur $[0,1]$ donc $Y = \frac{1}{X}$ prend ses valeurs sur $[1, +\infty[$. Ainsi $F_Y(y) = 0$ si $y \leq 1$. Soit $y \geq 1$, alors :

$$\begin{aligned} F_Y(y) &= P(\frac{1}{X} \leq y) \\ &= P(\frac{1}{y} \leq X) \quad \text{car } \frac{1}{y} > 0 \,\text{et } X \text{ ne prend que des valeurs positives,} \\ &= 1 - F_X(\frac{1}{y}) \end{aligned}$$

Or :

$$F_X(x) = \int_0^x 3t^2\,dt = x^3 \quad \text{pour } x \in [0,1]\,,$$

donc :

$$F_Y(y) = 1 - \frac{1}{y^3} \quad \text{pour } y \geq 1 \text{ (et 0 sinon).}$$

On vérifie que la continuité est assurée pour $y = 1$ car l'expression ci-dessus vaut 0 pour $y = 1$. Par dérivation on a :

$$f_Y(y) = \frac{3}{y^4} \quad \text{pour } y \geq 1 \text{ (et 0 sinon),}$$

d'où :

$$\begin{aligned} E(Y) &= \int_1^{+\infty} y\left(\frac{3}{y^4}\right) dy = 3\int_1^{+\infty} \frac{1}{y^3}\,dy \\ &= 3\left[\frac{y^{-2}}{-2}\right]_1^{+\infty} = \frac{3}{2}. \end{aligned}$$

Chapitre 3 : Couples et n-uplets de variables aléatoires

Exercice 3.1

On a :

$$\begin{aligned}
E(X) &= 1 \times 0,35 + 2 \times 0,45 + 3 \times 0,20 = 1,85 \\
E(Y) &= 1 \times 0,48 + 2 \times 0,33 + 3 \times 0,19 = 1,71 \\
E(X^2) &= 1^2 \times 0,35 + 2^2 \times 0,45 + 3^2 \times 0,20 = 3,95 \\
E(Y^2) &= 1^2 \times 0,48 + 2^2 \times 0,33 + 3^2 \times 0,19 = 3,51 \\
V(X) &= 3,95 - (1,85)^2 = 0,5275 \quad V(Y) = 3,51 - (1,71)^2 = 0,5859 \\
E(XY) &= 1 \times 1 \times 0,22 + 1 \times 2 \times 0,11 + \cdots + 3 \times 3 \times 0,07 = 3,33 \\
cov(X,Y) &= 3,33 - 1,85 \times 1,71 = 0,1665
\end{aligned}$$

et finalement :

$$corr(X,Y) = \frac{0,1665}{\sqrt{0,5275 \times 0,5859}} = 0,2995.$$

Exercice 3.2

Soit X et Y les deux variables aléatoires. On a, quelles qu'elles soient :

$$\begin{aligned}
cov(X+Y, X-Y) &= cov(X+Y, X) - cov(X+Y, Y) \\
&= cov(X,X) + cov(Y,X) - cov(X,Y) - cov(Y,Y) \\
&= V(X) - V(Y).
\end{aligned}$$

Il suffit que les deux variables soient de même loi pour que la covariance soit nulle. L'indépendance n'est pas nécessaire.

Exercice 3.3

L'événement $(X + Y = 0)$ équivaut à $(X = 0, Y = 0)$ – c'est-à-dire $(X = 0) \cap (Y = 0)$ – dont, par l'indépendance, la probabilité vaut $(1-p)(1-p)$.

De même $(X+Y=0)$ équivaut à : $(X=0, Y=1)$ ou $(X=1, Y=0)$ soit une probabilité totale de $2 \times p(1-p)$, etc.

La loi de $X+Y$ est définie par :

valeurs possibles	0	1	2
avec probabilités	$(1-p)^2$	$2p(1-p)$	p^2

On établit de la même façon la loi de $X-Y$:

valeurs possibles	-1	0	1
avec probabilités	$p(1-p)$	$p^2+(1-p)^2$	$p(1-p)$

L'événement $(X+Y=0,\ X-Y=0)$ ne se réalise que si et seulement si $(X=0, Y=0)$ (puisque le système des deux équations $x+y=0$ et $x-y=0$ n'a qu'une seule solution : $x=0$ et $y=0$). Par conséquent :

$$P(X+Y=0,\ X-Y=0) = P(X=0, Y=0) = (1-p)^2,$$

alors que $P(X+Y=0)P(\,X-Y=0) = (1-p)^2[p^2+(1-p)^2]$, ce qui est différent dans les cas non dégénérés (i.e. $p \neq 0$ et $p \neq 1$). On en déduit donc que $X+Y$ et $X-Y$ ne sont pas deux v.a. indépendantes. Toutefois $cov(X+Y, X-Y) = 0$ comme il a été montré dans l'exercice 3.2 ci-dessus, ce qui illustre le fait qu'une corrélation linéaire nulle n'implique pas nécessairement l'indépendance.

Exercice 3.4

Calculons $P(Z \leq z | X = x)$.

$$\begin{aligned} P(Z \leq z|X=x) &= P(X+Y \leq z|X=x) = P(x+Y \leq z|X=x) \\ &= P(Y \leq z-x|X=x) \\ &= P(Y \leq z-x) \text{ car } X \text{ et } Y \text{ sont indépendantes.} \end{aligned}$$

Donc $F_{Z|X=x}(z) = F_Y(z-x)$ et, en dérivant par rapport à z, $f_{Z|X=x}(z) = f_Y(z-x)$.

On a montré en fin de section 3.2 que :

$$f_{Z|X=x}(z) = \frac{f_{X,Z}(x,z)}{f_X(x)},$$

d'où :

$$f_{X,Z}(x,z) = f_Y(z-x) f_X(x)$$

et (voir aussi section 3.2) :

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Z}(x,z)\,dx = \int_{-\infty}^{+\infty} f_Y(z-x) f_X(x)\,dx\,.$$

Notons que par le changement de variable $y = z - x$ cette densité s'écrit également $\int_{-\infty}^{+\infty} f_Y(y) f_X(z-y)\,dy$

Pour $T = X - Y$ on se ramène au cas précédent en posant $U = -Y$, d'où $T = X + U$. La loi de U est donnée par :

$$F_U(y) = P(U \le y) = P(-Y \le y) = P(-y \le Y) = 1 - F_Y(-y)\,,$$

d'où $f_U(y) = f_Y(-y)$. En appliquant la formule plus haut, on obtient :

$$f_T(t) = \int_{-\infty}^{+\infty} f_U(t-x) f_X(x)\,dx = \int_{-\infty}^{+\infty} f_Y(x-t) f_X(x)\,dx\,,$$

ce qui s'écrit encore, par changement de variable $y = x - t$:

$$\int_{-\infty}^{+\infty} f_Y(y) f_X(t+y)\,dy\,.$$

Exercice 3.5

Comme $f_X(x) = 1$ pour $0 \le x \le 1$ et $f_Y(y) = 1$ pour $0 \le y \le 1$, par l'indépendance on a $f_{X,Y}(x,y) = 1$ dans le carré du plan $\{(x,y)\,|\,0 \le x \le 1,\ 0 \le y \le 1\}$ et 0 ailleurs. Ainsi la probabilité de toute région à l'intérieur du carré est égale à son aire. Posons $Z = X + Y$ et calculons sa fonction de répartition $P(Z \le z)$. On voit sur la figure 3.1 qu'il faut distinguer le cas $z \le 1$ du cas $z > 1$.

1) Pour $z \in [0,1]$
L'événement $(X + Y \le z)$ correspond alors aux réalisations (x,y) du couple (X,Y) appartenant au domaine A indiqué sur la figure. Son aire étant égale à $z^2/2$ on a $P(Z \le z) = z^2/2$ si $0 \le z \le 1$.

2) Pour $z \in [1,2]$
L'événement $(X + Y \le z)$ correspond alors aux réalisations (x,y) du couple (X,Y) appartenant au carré, à l'exclusion du domaine B indiqué sur la figure dont l'aire est $(2-z)^2/2$. D'où $P(Z \le z) = 1 - (2-z)^2/2$ si $1 < z \le 2$. Donc :

$$F_{X+Y}(z) = \begin{cases} 0 & \text{si } z \le 0 \\ z^2/2 & \text{si } 0 \le z \le 1 \\ (2-z)^2/2 & \text{si } 1 \le z \le 2 \\ 1 & \text{si } 2 \le z \end{cases}.$$

On vérifie qu'il y a continuité entre les différents morceaux de la fonction de répartition.

Dérivons pour décrire la densité :

$$f_{X+Y}(z) = \begin{cases} z & \text{si } 0 \leq z \leq 1 \\ 2-z & \text{si } 1 \leq z \leq 2 \\ 0 & \text{sinon} \end{cases}$$

dont le graphe fait apparaître un triangle dont la base est le segment $[0, 2]$ sur l'axe des abcisses. Cette loi est naturellement appelée *loi triangulaire*.

Note : Pour le calcul de la densité on aurait pu appliquer directement le résultat de l'exercice 3.4 mais en étant attentif aux bornes de l'intégrale. En premier lieu $\int_{-\infty}^{+\infty} f_Y(z-x) f_X(x)\, dx$ se ramène à $\int_{z-1}^{z} f_X(x)\, dx$ puisque $f_Y(z-x) = 1$ quand $0 < z - x < 1$, soit $z - 1 < x < z$. Pour calculer cette dernière intégrale, il faut distinguer le cas $0 < z < 1$ où elle vaut $\int_0^z 1\, dx = z$, et le cas $1 < z < 2$ où elle vaut $\int_{z-1}^1 1\, dx = 2 - z$.

Exercice 3.6

La surface étant XY, on a $E(XY) = E(X)E(Y) = \mu_X \mu_Y$ (voir la proposition 3.7).

De plus, $V(XY) = E((XY)^2) - [E(XY)]^2$. Or $E((XY)^2) = E(X^2Y^2) = E(X^2)E(Y^2)$. En effet, comme X et Y sont indépendantes, X^2 et Y^2 le sont aussi (proposition 3.4). Puisque $E(X^2) = \sigma_X^2 + \mu_X^2$ et $E(Y^2) = \sigma_Y^2 + \mu_Y^2$, on a :

$$V(XY) = (\sigma_X^2 + \mu_X^2)(\sigma_Y^2 + \mu_Y^2) - (\mu_X \mu_Y)^2 = \sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_Y^2 \sigma_X^2.$$

Exercice 3.7

La v.a. X ayant pour fonction de probabilité :

$$p(x) = \begin{cases} p & \text{si } x = 1 \\ 1-p & \text{si } x = -1 \end{cases},$$

on a :

$$\begin{aligned} E(X) &= p + (-1)(1-p) = 2p - 1 \\ E(X^2) &= 1^2 p + (-1)^2 (1-p) = 1 \\ V(X) &= E(X^2) - [E(X)]^2 = 1 - (2p-1)^2 = 4p(1-p). \end{aligned}$$

Soit maintenant n étapes successives de déplacements $X_1, X_2, \cdots, X_n$, ces v.a. étant indépendantes et chacune suivant la loi ci-dessus. La position résultante étant $Y = \sum_{i=1}^{n} X_i$, on a :

$$E(Y) = nE(X) = n(2p-1)$$
$$V(Y) = nV(X) = 4np(1-p).$$

De toute évidence, si $p > 1/2$ alors $E(Y) > 0$ et, inversement, si $p < 1/2$ alors $E(Y) < 0$.

Exercice 3.8

Appliquons avec $p = 2$ la formule générale de la densité gaussienne d'un $p-$vecteur gaussien :

$$f_{\mathbf{X}}(x_1, x_2, \cdots, x_p) = \frac{1}{(2\pi)^{p/2}(\det \mathbf{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \ \mathbf{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

La matrice des variances-covariances est :

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

L'inverse d'une $2\times 2-$matrice inversible $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ étant :

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad \text{où } \det A = ad - bc\,,$$

on a $\det \Sigma = \sigma_X^2\sigma_Y^2(1-\rho^2)$ et :

$$\Sigma^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}.$$

L'expression dans l'exponentielle se réduit ici à $A = -\frac{1}{2}(x,y)\ \mathbf{\Sigma}^{-1}\begin{pmatrix} x \\ y \end{pmatrix}$.

Comme $(x,y)\begin{pmatrix} a & c \\ c & b \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2cxy + by^2$ on a :

$$A = -\frac{1}{2\sigma_X^2\sigma_Y^2(1-\rho^2)}(\sigma_Y^2x^2 - 2\rho\sigma_X\sigma_Y xy + \sigma_X^2y^2)$$
$$= -\frac{1}{2(1-\rho^2)}(\frac{x^2}{\sigma_X^2} - 2\rho\frac{xy}{\sigma_X\sigma_Y} + \frac{y^2}{\sigma_Y^2})$$

Par ailleurs, la constante devant l'exponentielle est bien $\frac{1}{2\pi(\det \mathbf{\Sigma})^{\frac{1}{2}}}$.

Chapitre 4 : Les lois de probabilités usuelles

Exercice 4.1

La marche aléatoire est décrite dans l'exercice 3.7.

Notons que lorsque X prend respectivement les valeurs 1 et -1, $Y = \frac{1}{2}(X+1)$ prend les valeurs 1 et 0. X prenant la valeur 1 avec probabilité p, Y suit une loi de Bernoulli $\mathcal{B}(p)$.

La v.a. de la marche aléatoire après n pas est $T = \sum_{i=1}^{n} X_i$.
Comme $X_i = 2Y_i - 1$, avec $Y_i \rightsquigarrow \mathcal{B}(p)$, on a :

$$T = 2\sum_{i=1}^{n} Y_i - n = 2S - n$$

où $S = \sum_{i=1}^{n} Y_i$ suit une loi binomiale $\mathcal{B}(n,p)$:

$$P(S=k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ pour } k = 0, 1, 2, \cdots, n.$$

Comme l'événement $(S = k)$ équivaut à $(T = 2k - n)$ on a, en posant $t = 2k - n$, la loi suivante pour T :

$$P(T=t) = \binom{n}{\frac{n+t}{2}} p^{\frac{n+t}{2}} (1-p)^{\frac{n-t}{2}} \text{ pour } t = -n, -n+2, \cdots, n-2, n.$$

Exercice 4.2

Soit X qui suit une loi binomiale négative $\mathcal{BN}(r,p)$, d'où (voir section 4.1.4) :

$$P(X=x) = \binom{r+x-1}{x} p^r (1-p)^x \text{ pour } x \in \mathbb{N}.$$

Sa fonction génératrice des moments est (section 2.5) :

$$\Psi_X(t) = E(e^{tX}) = \sum_{x\in\mathbb{N}} e^{tx} \binom{r+x-1}{x} p^r (1-p)^x$$
$$= p^r \sum_{x\in\mathbb{N}} \binom{r+x-1}{x} [(1-p)e^t]^x.$$

Pour calculer la somme ci-dessus, posons $1-q=(1-p)e^t$ et notons qu'en vertu de la fonction de probabilité d'une loi $\mathcal{BN}(r,q)$ on a :

$$\sum_{x\in\mathbb{N}}\binom{r+x-1}{x}q^r(1-q)^x=1$$

et :

$$\sum_{x\in\mathbb{N}}\binom{r+x-1}{x}(1-q)^x=\frac{1}{q^r}.$$

Notons que ceci est vrai si $0<1-q<1$ et que, de toute façon, les sommes ci-dessus sont divergentes si $1-q>1$ puisque $(1-q)^x\to+\infty$ quand $x\to+\infty$. Il est donc nécessaire que $0<(1-p)e^t<1$ soit $0<e^t<\frac{1}{1-p}$ ou $t<\ln\frac{1}{1-p}$ ou $t<-\ln(1-p)$

Finalement, en revenant à la dernière écriture de$\Psi_X(t)$, on a :

$$\begin{aligned}\Psi_X(t)&=p^r\frac{1}{q^r}\\&=\frac{p^r}{[1-(1-p)e^t]^r}\end{aligned}$$

qui est bien l'expression donnée en section 4.1.4.

Pour obtenir $E(X)$, dérivons par rapport à p chaque terme de l'égalité :

$$\sum_{x\in\mathbb{N}}\binom{r+x-1}{x}p^r(1-p)^x=1$$

ce qui donne :

$$r\sum_{x\in\mathbb{N}}\binom{r+x-1}{x}p^{r-1}(1-p)^x-\sum_{x\in\mathbb{N}}x\binom{r+x-1}{x}p^r(1-p)^{x-1}=0,$$

soit en multipliant par p :

$$\begin{aligned}r-\frac{p}{1-p}\sum_{x\in\mathbb{N}}x\binom{r+x-1}{x}p^r(1-p)^x&=0\\r-\frac{p}{1-p}E(X)&=0\end{aligned}$$

d'où $E(X)=\frac{r(1-p)}{p}$.

On peut évidemment arriver à ce résultat en calculant $\Psi_X^{'}(0)$ comme indiqué en proposition 2.2.

Exercice 4.3

On posera $q = 1 - p$ pour simplifier les écritures.

Notons que, par interchangeabilité des notions de succès et échec, $n-x \to \infty$ pour les valeurs d'intérêt de $n - x$.

On a :

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

En appliquant la formule de Stirling pour $n!$, $x!$ et $(n - x)!$ on obtient :

$$P(X = x) \sim \frac{1}{\sqrt{2\pi}} \left[\frac{n}{x(n-x)}\right]^{1/2} \frac{n^x n^{n-x}}{x^x (n-x)^{n-x}} p^x q^{n-x}$$

$$P(X = x) \sim \frac{1}{\sqrt{2\pi}} \left[\frac{1}{n\frac{x}{n}(1-\frac{x}{n})}\right]^{1/2} \left[\left(\frac{x}{np}\right)^x \left(\frac{n-x}{nq}\right)^{n-x}\right]^{-1},$$

soit, en posant $u = x - np$ (donc $\frac{u}{n} \to 0$ puisque $\frac{x}{n} \to p$) :

$$P(X = x) \sim \frac{1}{\sqrt{2\pi npq}} \left[\left(1 + \frac{u}{np}\right)^{u+np} \left(1 - \frac{u}{nq}\right)^{u-nq}\right]^{-1}$$

Le logarithme de l'expression entre parenthèses est égal à :

$$(u + np) \ln(1 + \frac{u}{np}) + (u - nq) \ln(1 - \frac{u}{nq})$$

et équivalent à :

$$(u + np) \left[\frac{u}{np} - \frac{u^2}{2n^2p^2}\right] + (u - nq) \left[-\frac{u}{nq} - \frac{u^2}{2n^2q^2}\right] = \frac{u^2}{2npq}$$

donc :

$$P(X = x) \sim \frac{1}{\sqrt{2\pi npq}} \exp\left\{-\frac{u^2}{2npq}\right\} = \frac{1}{\sqrt{2\pi npq}} \exp\left\{-\frac{(x-np)^2}{2npq}\right\}$$

qui est la fonction de la densité de U en x.

Montrons maintenant que cette expression est équivalente à :

$$P(x - \frac{1}{2} < U < x + \frac{1}{2}) = F_U(x + \frac{1}{2}) - F_U(x - \frac{1}{2}).$$

Par la formule des accroissements finis on a :

$$F_U(x + \frac{1}{2}) - F_U(x - \frac{1}{2}) = f_U(x + h)$$

où $h \in (-\frac{1}{2}, +\frac{1}{2})$ et f_U est la fonction de densité de U, donc :

$$f_U(x+h) = \frac{1}{\sqrt{2\pi npq}} \exp\left\{-\frac{(x+h-np)^2}{2npq}\right\}.$$

Mais $f_U(x+h) \sim f_U(x)$ car :

$$\frac{f_U(x+h)}{f_U(x)} = \exp\left\{\frac{-(x+h-np)^2 + (x-np)^2}{2npq}\right\} = \exp\left\{-\frac{h}{2pq}\left[2(\frac{x}{n} - p) + \frac{h}{n}\right]\right\}$$

tend vers 1 quand $n \to \infty$, ce qui montre que $P(X = x) \sim P(x - \frac{1}{2} < U < x + \frac{1}{2})$. Cela constitue une approximation d'une loi binomiale par une loi de Gauss quand n est grand (voir section 5.8.3).

Annexe :
Justifions sommairement le fait que les valeurs utiles de x tendent vers l'infini en recourant à l'inégalité de Tchebichev introduite dans l'exercice 5.9. Selon cette inégalité, pour toute v.a. X ayant une variance et pour tout $k > 0$, on a :

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

soit, ici, en choisissant $k = \sqrt{np}$:

$$P(|X - np| < np\sqrt{q}) \geq 1 - \frac{1}{np}$$

$$P\left(np(1-\sqrt{q}) < X < np(1+\sqrt{q})\right) \geq 1 - \frac{1}{np}$$

et donc l'ensemble des valeurs inférieures à $np(1-\sqrt{q})$, qui tend vers l'infini, reçoit une probabilité tendant vers 0.

Par ailleurs, le fait que pour les valeurs d'intérêt de x on ait $\frac{x}{n} \to p$ découle de la loi des grands nombres (section 5.8.2).

Exercice 4.4

Cet exercice demande une démonstration directe de la propriété établie par la fonction génératrice en section 4.1.7.

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \sim \frac{(np)^x}{x!}(1-p)^{n-x}$$

puisque $\frac{n!}{(n-x)!} \sim n^x$. En posant $np = \lambda$ (constant quand $n \to \infty$), on a $(1-p)^{n-x} = (1-\frac{\lambda}{n})^{n-x}$, lequel tend vers $e^{-\lambda}$ quand $n \to \infty$, et donc $P(X = x)$ tend vers $\frac{\lambda^x e^{-\lambda}}{x!}$ qui est la probabilité de la loi $\mathcal{P}(\lambda)$ en x.

Exercice 4.5

Pour cette loi hypergéométrique, on a :

$$P(X=x)=\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}.$$

et notons que $N\to\infty$ et $\frac{M}{N}\to p\neq 0$ implique que $M\to\infty$ et $N-M\to\infty$. En utilisant le fait que $\frac{s!}{(s-u)!}\sim s^u$, ou $\binom{s}{u}\sim\frac{s^u}{u!}$, quand s tend vers l'infini et u est fixé (s et u entiers positifs), on peut écrire :

$$\begin{aligned}P(X=x)&\sim\frac{M^x}{x!}\frac{(N-M)^{n-x}}{(n-x)!}\frac{n!}{N^n}\\&\sim\binom{n}{x}\left(\frac{M}{N}\right)^x\left(1-\frac{M}{N}\right)^{n-x}\end{aligned}$$

et, ainsi, $P(X=x)$ tend vers $\binom{n}{x}p^x\,(1-p)^{n-x}$.

Exercice 4.6

Calculons :

$$\begin{aligned}P(X_1=k\,|\,X_1+X_2=n)&=\frac{P((X_1=k)\cap(X_2=n-k))}{P(X_1+X_2=n)}\\&=\frac{\frac{e^{-\lambda_1}\lambda_1^k}{k!}\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}}{\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}}=\binom{n}{k}\left[\frac{\lambda_1}{\lambda_1+\lambda_2}\right]^k\left[\frac{\lambda_2}{\lambda_1+\lambda_2}\right]^{n-k}\end{aligned}$$

qui est la probabilité pour k de la loi binomiale de paramètres n et $\frac{\lambda_1}{\lambda_1+\lambda_2}$.

Exercice 4.7

Reprenant la définition de la loi $\mathcal{G}(p)$, on a :

$$\begin{aligned}P(X>n)&=\sum_{x\geq n+1}p(1-p)^x=p(1-p)^{n+1}\left[1+(1-p)+(1-p)^2+\cdots\right]\\&=p(1-p)^{n+1}\frac{1}{1-(1-p)}=(1-p)^{n+1}.\end{aligned}$$

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

qui est indépendant de n et égal à $P(X \geq k)$. Dans la modélisation en temps discret d'une durée de vie cela signifie que la probabilité de vivre encore k unités de temps au-delà d'un temps donné reste constante.

Exercice 4.8

Reprenant la fonction de répartition de la loi $\mathcal{U}[0,1]$: $P(X \leq u) = u$ si $u \in [0,1]$ et vaut 0 si $u < 0$ ou 1 si $u > 1$.

$$\begin{aligned}P(Y \leq y) &= P((b-a)X + a \leq y) \\ &= P(X \leq \frac{y-a}{b-a})\end{aligned}$$

qui vaut $\frac{y-a}{b-a}$ si $\frac{y-a}{b-a} \in [0,1]$, soit $a \leq y \leq b$, 0 si $\frac{y-a}{b-a} < 0$, soit $y \leq a$, et 1 si $\frac{y-a}{b-a} > 1$, soit $y > b$. Cela correspond bien à la fonction de répartition de la loi $\mathcal{U}[a,b]$ définie en section 4.2.1.

Exercice 4.9

En section 4.2.3 la loi $\Gamma(r,\lambda)$ a été définie pour r entier ($r > 0$). Pour r, réel positif, on vérifie aisément (par changement de variable $u = \lambda x$) que la fonction de x :

$$\frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$$

est bien une fonction de densité, de par la définition même de $\Gamma(r)$.

Si elle existe, la fonction génératrice de $X \rightsquigarrow \Gamma(r,\lambda)$ est :

$$\begin{aligned}\Psi(t) &= \int_0^{+\infty} e^{tx} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} dx \\ &= \frac{\lambda^r}{(\lambda-t)^r} \int_0^{+\infty} \frac{(\lambda-t)^r}{\Gamma(r)} x^{r-1} e^{-(\lambda-t)x} dx \quad \text{si } t < \lambda.\end{aligned}$$

Posons $u = (\lambda - t)x$, alors :

$$\Psi(t) = \left(\frac{\lambda}{\lambda - t}\right)^r \int_0^{+\infty} \frac{1}{\Gamma(r)} u^{r-1} e^{-u} du.$$

Comme la fonction intégrée est la densité de la loi $\Gamma(r, 1)$, on a $\Psi(t) = \left(\frac{\lambda}{\lambda - t}\right)^r$.

Si $t > \lambda$ on pose $u = (t - \lambda)$ pour constater que la fonction à intégrer contient e^u en lieu et place de e^{-u} et l'intégrale est divergente.

Pour calculer $E(X)$, dérivons $\Psi(t)$:

$$\Psi'(t) = \frac{r\lambda^r}{(\lambda - t)^{r+1}} \quad \Longrightarrow \quad E(X) = \Psi'(0) = \frac{r}{\lambda},$$

puis :

$$\Psi''(t) = \frac{r(r+1)\lambda^r}{(\lambda - t)^{r+2}} \quad \Longrightarrow \quad E(X^2) = \Psi''(0) = \frac{r(r+1)}{\lambda^2}$$

et $V(X) = \frac{r(r+1)}{\lambda^2} - \left(\frac{r}{\lambda}\right)^2 = \frac{r}{\lambda^2}$. Notons que les résultats obtenus sont identiques à ceux de la section 4.2.3.

Exercice 4.10

Soit la v.a. T dénotant le temps entre une occurrence et la r-ième suivante dans un processus de Poisson d'intensité λ. Alors $T \rightsquigarrow \Gamma(r, \lambda)$ (voir section 4.2.3). L'évènement $(T \geq x)$ est identique au fait qu'il y ait moins de r occurrences dans un intervalle de temps de longueur x et, donc :

$$P(T \geq x) = \sum_{k=0}^{r-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

$$F_T(x) = 1 - \sum_{k=0}^{r-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

où $F_T(x)$ est la fonction de répartition de T. En dérivant, on obtient sa fonction de densité :

$$\begin{aligned} f_T(x) &= \lambda \sum_{k=0}^{r-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!} - \lambda \sum_{k=1}^{r-1} \frac{e^{-\lambda x}(\lambda x)^{k-1}}{(k-1)!} \\ &= \lambda \sum_{k=0}^{r-1} \frac{e^{-\lambda x}(\lambda x)^k}{k!} - \lambda \sum_{k=0}^{r-2} \frac{e^{-\lambda x}(\lambda x)^k}{k!} \\ &= \frac{\lambda^r e^{-\lambda x} x^{r-1}}{(r-1)!} \quad (x > 0) \end{aligned}$$

qui est bien l'expression donnée en section 4.2.3.

Exercice 4.11

Soit $X \rightsquigarrow \Gamma(r, \lambda)$, alors, avec des notations évidentes :

$$F_{rX}(x) = P(rX \leq x) = P(X \leq \frac{x}{r}) = F_X(\frac{x}{r}),$$
$$f_{rX}(x) = \frac{1}{r} f_X(\frac{x}{r}) = \frac{1}{r} \frac{\lambda^r e^{-\frac{\lambda x}{r}} x^{r-1}}{(r-1)! r^{r-1}} = \frac{(\frac{\lambda}{r})^r e^{-\frac{\lambda}{r}x} x^{r-1}}{(r-1)!}$$

qui est la densité de la loi $\Gamma(r, \frac{\lambda}{r})$. De la même façon, $F_{\lambda X}(x) = F_X(\frac{x}{\lambda})$, $f_{\lambda X}(x) = \frac{1}{\lambda} f_X(\frac{x}{\lambda}) = \frac{1}{\lambda} \frac{\lambda^r e^{-x} x^{r-1}}{(r-1)! \lambda^{r-1}} = \frac{e^{-x} x^{r-1}}{(r-1)!}$, densité de la loi $\Gamma(r, 1)$.

Exercice 4.12

Pour la v.a. X de loi de Pareto de seuil $a = 1$, on a $F_X(x) = 1 - x^{-\theta}$ pour $x \geq 1$ et 0 pour $x < 1$, où $\theta > 0$. Pour $Y = \ln(X)$, on a :

$F_Y(y) = P(Y \leq y) = P(e^Y \leq e^y) = P(X \leq e^y) = 1 - e^{-\theta y}$.

Comme $F_X(x)$ est nulle pour $x \leq 1$, $F_Y(y)$ vaut 0 pour $y \leq 0$, ce qui restitue la loi $\mathcal{E}(\theta)$.

Exercice 4.13

On a un processus de Poisson avec un nombre moyen d'arrivée par seconde $\lambda = 1/30$. Soit X le temps écoulé entre le départ du guichet de la 1-ère personne et celui de la sixième, alors $X \rightsquigarrow \Gamma(5, \lambda)$. On cherche à calculer $P(X \leq 30)$, soit $P(X \leq \frac{1}{\lambda})$ ou $P(\lambda X \leq 1)$. Comme $\lambda X \rightsquigarrow \Gamma(5, 1)$ selon le résultat de l'exercice 4.11, cette probabilité vaut $\int_0^1 \frac{1}{4!} x^4 e^{-x} dx$. Posons $I_n = \int_0^1 x^n e^{-x} dx$, alors, en intégrant par parties :

$$I_n = -\int_0^1 x^n d(e^{-x}) = -[x^n e^{-x}]_0^1 + n \int_0^1 x^{n-1} e^{-x} dx.$$

D'où la relation de récurrence $I_n = nI_{n-1} - e^{-1}$. En partant de $I_0 = \int_0^1 e^{-x} dx = 1 - e^{-1}$, on trouve $I_1 = 1 - 2e^{-1}$, $I_2 = 2 - 5e^{-1}$, $I_3 = 6 - 16e^{-1}$ et $I_4 = 24 - 65e^{-1}$. Finalement, en divisant I_4 par 4 ! on obtient une probabilité d'environ 0,0366. On peut accéder directement au résultat avec EXCEL en évaluant $LOI.GAMMA(30\,;\,5\,;\,30\,;\,VRAI)$.

Exercice 4.14

La v.a. $Y = X - 1$ prend, respectivement, les valeurs 0 et 1 avec probabilités 0,7 et 0,3 et suit donc une loi de Bernoulli $\mathcal{B}(0,3)$. Soit $Y_1, Y_2, \cdots, Y_{20}$ des variables aléatoires i.i.d. de cette même loi, alors $\sum_{i=1}^{20} Y_i$ suit une loi binomiale $\mathcal{B}(20\,;\,0,3)$. Le parking aura une capacité suffisante si l'événement $(\sum_{i=1}^{20} Y_i + 20 \leq 29)$, soit $(\sum_{i=1}^{20} Y_i \leq 9)$, se réalise. Par EXCEL on obtient une probabilité 0,952 pour l'expression $LOI.BINOMIALE(9\,;\,20\,;\,0,3\,;\,VRAI)$. Alternativement, par l'approximation gaussienne introduite en section 5.8.3, on calcule $P(S \leq 9,5)$ pour $S \rightsquigarrow \mathcal{N}(20 \times 0,3\,;\,20 \times 0,3 \times 0,7)$, soit $P(Z \leq \frac{9,5-6}{\sqrt{4,2}}) \simeq P(Z \leq 1,71)$ pour $Z \rightsquigarrow \mathcal{N}(0\,;1)$, ce qui vaut 0,956. On notera que l'approximation est satisfaisante dans la mesure où $20 \times 0,3 \geq 5$ et $20 \times 0,7 \geq 5$.

Exercice 4.15

Selon la définition donnée en section 4.2.5, $X \rightsquigarrow L\mathcal{N}(\mu\,,\sigma^2)$ si $Y = \ln X \rightsquigarrow \mathcal{N}(\mu\,,\sigma^2)$. Ainsi μ et σ^2 sont la moyenne et la variance non pas de X mais de son logarithme. Pour trouver μ et σ^2, sachant que $E(X) = a$ et $V(X) = b$, il faut résoudre les équations suivantes obtenues à partir des expression s de $E(X)$ et de $V(X)$ établies en section 4.2.5 :

$$\left\{ \begin{array}{c} e^{\mu+\frac{1}{2}\sigma^2} = a \\ e^{2\mu+\sigma^2}(e^{\sigma^2}-1) = b \end{array} \right. \Leftrightarrow \left\{ \begin{array}{c} \mu + \frac{1}{2}\sigma^2 = \ln a \\ e^{\sigma^2} - 1 = \frac{b}{a^2} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{c} \sigma^2 = \ln(1+\frac{b}{a^2}) \\ \mu = \ln a - \frac{1}{2}\ln(1+\frac{b}{a^2}) \end{array} \right. .$$

Ici $a = 70$ et $b = (12)^2$ ce qui donne $\mu = 4,234$ et $\sigma^2 = 0,02896$.

Dès lors, on peut calculer des probabilités selon le modèle retenu, par exemple $P(X \leq 80) = P(Y \leq \ln 80) = P(Z \leq \frac{\ln 80 - 4,234}{\sqrt{0,02896}}) = P(Z \leq 0,870)$, où $Z \rightsquigarrow \mathcal{N}(0\,;1)$, soit une probabilité de 0,808.

Chapitre 5 : Lois fondamentales de l'échantillonnage

Exercice 5.1

La v.a. $T = \sum_{i=1}^{n} X_i$ est de loi $\Gamma(n, \lambda)$ (voir section 4.2.3). Soit $F_{\overline{X}}(x)$ la fonction de répartition de $\overline{X}$:

$$F_{\overline{X}}(x) = P(\overline{X} \leq x) = P(T \leq nx) = F_T(nx).$$

Passons aux fonctions de densité :

$$f_{\overline{X}}(x) = nf_T(nx) = \frac{n\lambda^n}{(n-1)!}(nx)^{n-1}e^{-\lambda(nx)} = \frac{(n\lambda)^n}{(n-1)!}x^{n-1}e^{-(\lambda n)x}$$

qui est la densité de la loi $\Gamma(n, n\lambda)$. Note : On peut aussi utiliser les fonctions génératrices sur le modèle de l'exercice suivant.

Exercice 5.2

Pour $T = \sum_{i=1}^{n} X_i$, on a la fonction génératrice $\Psi_T(t) = [\Psi_X(t)]^n$ où $\Psi_X(t)$ est la fonction génératrice de la loi $\Gamma(r, \lambda)$ (voir proposition 3.12). D'où $\Psi_T(t) = \left(\frac{\lambda}{\lambda - t}\right)^{nr}$. Or :

$$\Psi_{\overline{X}}(t) = \Psi_{\frac{T}{n}}(t) = E(e^{\frac{T}{n}t}) = E(e^{T\frac{t}{n}}) = \Psi_T(\frac{t}{n}) = \left(\frac{\lambda}{\lambda - \frac{t}{n}}\right)^{nr} = \left(\frac{n\lambda}{n\lambda - t}\right)^{nr}$$

qui est la fonction génératrice de la loi $\Gamma(nr, n\lambda)$.

Exercice 5.3

Calculons la fonction caractéristique (voir note 2.4) de la v.a. X qui suit la loi de Cauchy définie lors de l'exemple 2.1. :

$$\begin{aligned}\Phi_X(t) &= E(e^{itX}) = E(\cos tX) + iE(\sin tX) \\ &= \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{\cos tx}{1+x^2}dx + i\frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{\sin tx}{1+x^2}dx = \frac{2}{\pi}\int_{0}^{+\infty} \frac{\cos tx}{1+x^2}dx = e^{-|t|},\end{aligned}$$

d'où $\Phi_{\overline{X}}(t) = E\left(e^{\frac{1}{n}\sum X_i \cdot t}\right) = \prod_{i=1}^n \Phi_{X_i}(\frac{t}{n}) = \prod_{i=1}^n e^{-\frac{|t|}{n}} = e^{-|t|}$. La loi de Cauchy n'ayant pas de moyenne (ni a fortiori de variance), la loi des grands nombres ne s'applique pas.

Exercice 5.4

La statistique $(n-1)S^2 = \sum_{i=1}^n (X_i - \overline{X})^2$ peut s'écrire (voir la démonstration de la proposition 5.2) $\sum_{i=1}^n (X_i - \mu)^2 - n(\overline{X} - \mu)^2$ ou encore $\sum_{i=1}^n (X_i - \mu)^2 - \frac{1}{n}\left[\sum_{i=1}^n (X_i - \mu)\right]^2$. Soit $Z_i = X_i - \mu$. On a $E(Z_i) = 0$, $E(Z_i^2) = \sigma^2$ et $E(Z_i^4) = \mu'_4$, où σ^2 et μ'_4 sont, respectivement, les deuxième et quatrième moments centrés de la loi mère de l'échantillon aléatoire considéré. Les Z_i sont i.i.d. Par substitution et développement du carré d'une somme, on obtient :

$$(n-1)S^2 = \sum_{i=1}^n Z_i^2 - \frac{1}{n}\left[\sum_{i=1}^n Z_i^2 + 2\sum\sum_{i<j} Z_i Z_j\right] = \frac{n-1}{n}\sum_{i=1}^n Z_i^2 - \frac{2}{n}\sum\sum_{i<j} Z_i Z_j$$

$$\begin{aligned}(n-1)^2 E(\left[S^2\right]^2) &= E\left(\left[\frac{n-1}{n}\sum_{i=1}^n Z_i^2 - \frac{2}{n}\sum\sum_{i<j} Z_i Z_j\right]^2\right) \\ &= \left(\frac{n-1}{n}\right)^2 E\left(\sum_{k=1}^n Z_k^4 + 2\sum\sum_{i<j} Z_i^2 Z_j^2\right) \\ &\quad - \frac{4(n-1)}{n^2} E\left(\left[\sum_{k=1}^n Z_k^2\right]\left[\sum\sum_{i<j} Z_i Z_j\right]\right) + \frac{4}{n^2} E\left(\left[\sum\sum_{i<j} Z_i Z_j\right]^2\right)\end{aligned}$$

Le premier terme vaut $\left(\frac{n-1}{n}\right)^2 (n\mu'_4 + n(n-1)\sigma^4)$, le deuxième est nul car il contient soit des termes en $E(Z_i Z_j Z_k^2)$ avec $i \neq j \neq k$, soit des termes en $E(Z_j Z_k^3)$ avec $j \neq k$. Pour le dernier terme les produits croisés sont nuls car ils contiennent soit des termes en $E(Z_i^2 Z_j Z_k)$ avec $i \neq j \neq k$, soit en $E(Z_i Z_j Z_k Z_l)$ avec $i \neq j \neq k \neq l$. Il vaut donc $\frac{4}{n^2} E(\sum\sum_{i<j} Z_i^2 Z_j^2) = \frac{2(n-1)}{n}\sigma^4$. Finalement :

$$\begin{aligned}(n-1)^2 V(S^2) &= \frac{(n-1)^2}{n}\mu'_4 + \frac{(n-1)^3}{n}\sigma^4 + 2\frac{(n-1)}{n}\sigma^4 - (n-1)^2\left[E(S^2)\right]^2 \\ &= \frac{(n-1)^2}{n}\mu'_4 + \frac{(n-1)(3-n)}{n}\sigma^4\end{aligned}$$

$$V(S^2) = \frac{1}{n}(\mu'_4 - \frac{n-3}{n-1}\sigma^4).$$

Pour la loi de Gauss, $\mu'_4 = \frac{4!}{4\times 2!}\sigma^4 = 3\sigma^4$, donc $V(S^2) = \frac{1}{n}(3\sigma^4 - \frac{n-3}{n-1}\sigma^4) = \frac{2\sigma^4}{n-1}$. Note : On peut retrouver ce résultat sachant que $\frac{(n-1)S^2}{\sigma^2}$ suit une loi du

khi-deux à $n-1$ degrés de liberté dont la variance est égale à $2(n-1)$. D'où $\frac{(n-1)^2 V(S^2)}{\sigma^4} = 2(n-1)$ et $V(S^2) = \frac{2\sigma^4}{n-1}$.

Exercice 5.5

Quel que soit i les valeurs possibles pour X_i sont $a_1, a_2, \ldots, a_N$ et sa loi marginale est $P(X_i = a_j) = \frac{1}{N}$ pour $j = 1, 2, \ldots, N$.

La loi conjointe de X_i et X_k, $i \neq k$, est $P(X_i = a_j, X_k = a_l) = \frac{1}{N(N-1)}$, car il y a $N(N-1)$ arrangements pour les (a_j, a_l), tous équiprobables par symétrie. Il résulte de cela que :

$$E(X_i) = \frac{1}{N}\sum_{j=1}^{N} a_j = \mu, V(X_i) = \frac{1}{N}\sum_{j=1}^{N}(a_j - \mu)^2 = \sigma^2, E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

$$E(X_i X_k) = \frac{1}{N(N-1)}\sum_{j\neq l} a_j a_l = \frac{1}{N(N-1)}\left[(\sum_j a_j)^2 - \sum_j a_j^2\right],$$

d'où $E(X_i X_k) = \frac{1}{N(N-1)}\left[N^2\mu^2 - (N\sigma^2 + N\mu^2)\right]$ et $cov(X_i, X_k) = E(X_i X_k) - \mu^2 = -\frac{\sigma^2}{N-1}$.

Puis $V(\sum_{i=1}^n X_i) = \sum_{i=1}^n V(X_i) + 2\sum_{i<k} cov(X_i, X_k)$ et, comme il y a $\frac{n(n-1)}{2}$ termes de covariance ici,

$$V(\sum_{i=1}^{n} X_i) = n\sigma^2 - 2\frac{n(n-1)}{2}\frac{\sigma^2}{N-1} = n\sigma^2\frac{N-n}{N-1},$$

d'où $V(\overline{X}) = \frac{1}{n^2}V(\sum_{i=1}^n X_i) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$.

Exercice 5.6

Soit $Z_i = X_i - \mu$, alors $\bar{X} - \mu = \overline{Z}$, $E(\overline{Z}) = 0$ et $\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (Z_i - \overline{Z})^2 = \sum_{i=1}^n Z_i^2 - n\overline{Z}^2$. Donc :

$$cov(\bar{X}, S^2) = cov(\bar{X} - \mu, S^2) = cov\left(\overline{Z}, \frac{1}{n-1}\left[\sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2\right]\right)$$
$$= \frac{1}{n-1}E\left(\overline{Z}\left[\sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2\right]\right) = \frac{1}{n-1}\left[E\left(\overline{Z}\left[\sum_{i=1}^{n} Z_i^2\right]\right) - nE(\overline{Z}^3)\right],$$

or $E\left(\overline{Z}(\sum_{i=1}^n Z_i^2)\right) = \frac{1}{n}E((\sum_{j=1}^n Z_j)(\sum_{i=1}^n Z_i^2)) = \frac{1}{n}E(\sum_{i=1}^n Z_i^3) = \mu_3^{'}$
et $E(\overline{Z}^3) = \frac{1}{n^3}E([\sum_{i=1}^n Z_i]\left[\sum_{j=1}^n Z_j\right][\sum_{k=1}^n Z_k]) = \frac{1}{n^3}E(\sum_{i=1}^n Z_i^3) = \frac{1}{n^2}\mu_3^{'}$.
D'où :

$$cov(\bar{X}, S^2) = \frac{1}{n-1}\left[\mu_3^{'} - \frac{1}{n}\mu_3^{'}\right] = \frac{\mu_3^{'}}{n}.$$

Exercice 5.7

Soit $Z \rightsquigarrow \mathcal{N}(0\,;1)$ et, donc, $Z^2 \rightsquigarrow \chi^2(1)$ de fonction de répartition F_{Z^2} et de densité f_{Z^2} Alors :

$$F_{Z^2}(x) = P(Z^2 \leq x) = P(-\sqrt{x} < Z < \sqrt{x}) = 2\Phi(\sqrt{x}) - 1 \text{ si } x > 0 \text{ (0 sinon)}.$$

En dérivant :

$$f_{Z^2}(x) = \frac{2}{2\sqrt{x}}\Phi'(\sqrt{x}) = \frac{1}{\sqrt{x}\sqrt{2\pi}}e^{-\frac{x}{2}} \text{ si } x > 0 \text{ (0 sinon)}.$$

Exercice 5.8

La fonction génératrice de la loi $\chi^2(\nu)$ est $\Psi(t) = (1-2t)^{-\frac{\nu}{2}}$ (voir section 5.3). Sa moyenne est donc $\mu = \Psi'(0)$. Comme $\Psi'(t) = \nu(1-2t)^{-\frac{\nu}{2}-1}$ on a $\mu = \nu$. Le moment simple d'ordre 2 est $\mu_2 = \Psi''(0)$ avec $\Psi''(t) = \nu(\nu+2)(1-2t)^{-\frac{\nu}{2}-2}$, soit $\mu_2 = \nu(\nu+2)$. La variance est le moment centré d'ordre 2 : $\mu_2' = \mu_2 - \mu^2 = 2\nu$.

Exercice 5.9

1. $\int_{-\infty}^{\infty} g(x)f_X(x)dx \geq \int_A g(x)f_X(x)dx > k\int_A f_X(x)dx$.
   Donc $E(g(X)) > k\int_A f_X(x)dx$.

   Mais $\int_A f_X(x)dx = P(X \in A) = P(g(X) > k)$. D'où :

   $E(g(X)) \geq kP(g(X) > k)$.

2. En prenant $g(x) = (x-\mu)^2$ et en posant $\varepsilon^2 = k$, $\varepsilon > 0$, on a :

   $$\sigma^2 = E((X-\mu)^2) \geq \varepsilon^2 P((X-\mu)^2) \geq \varepsilon^2) = \varepsilon^2 P(|X-\mu| > \varepsilon)$$

   ce qui démontre l'inégalité de Tchebichev.

3. Posons $X = Y_n - Y$. Alors $E((Y_n - Y)^2) = E(X^2) \geq \varepsilon^2 P(|X| > \varepsilon) = \varepsilon^2 P(|Y_n - Y| > \varepsilon)$ pour tout $\varepsilon > 0$ donné.
$Y_n \xrightarrow{m.q.} Y$ équivaut, par définition, à $\lim\limits_{n\to\infty} E((Y_n - Y)^2) \to 0$. Donc $\lim\limits_{n\to\infty} P(|Y_n - Y| > \varepsilon) \to 0$ ou $\lim\limits_{n\to\infty} P(|Y_n - Y| < \varepsilon) \to 1$ ce qui définit $Y_n \xrightarrow{p} Y$.

Exercice 5.10

Soit $X_1, \cdots, X_n$ de loi mère de moyenne μ et de variance σ^2. Appliquons l'inégalité de Tchebichev à $\overline{X}_n$ de moyenne μ et variance $\frac{\sigma^2}{n}$:

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \quad \text{pour tout } \varepsilon > 0 \text{ fixé.}$$

$$\text{Donc } \lim_{n\to\infty} P(|\overline{X}_n - \mu| > \varepsilon) \to 0 \iff \lim_{n\to\infty} P(|\overline{X}_n - \mu| < \varepsilon) \to 1$$

soit, par définition, $\overline{X}_n \xrightarrow{p} \mu$ (convergence faible).

Exercice 5.11

La médiane de la loi mère est le nombre $M = F^{-1}(\frac{1}{2})$ où F est sa fonction de répartition. La fonction de répartition du maximum de l'échantillon $X_{(n)}$ est $[F(x)]^n$ en x (voir section 5.6), d'où :

$$P\left(X_{(n)} \leq F^{-1}(\frac{1}{2})\right) = [F(F^{-1}(\frac{1}{2}))]^n = \frac{1}{2^n} \quad \Longrightarrow \quad P(X_{(n)} > M) = 1 - \frac{1}{2^n}.$$

De même $P\left(X_{(n)} \leq F^{-1}(\frac{3}{4})\right) = (\frac{3}{4})^n$ et la probabilité que le maximum dépasse le troisième quartile de la loi mère est donc $1 - (\frac{3}{4})^n$. Notons que dans les deux cas la probabilité tend vers 1 quand la taille de l'échantillon croît vers l'infini.

Exercice 5.12

Pour la loi $\mathcal{U}[0\,,1]$ on a $F(x) = x$ pour $x \in [0\,,1]$. Pour le minimum $X_{(1)}$ de l'échantillon de taille n la fonction de répartition est :

$$\begin{aligned} F_{X_{(1)}}(x) &= 1 - [1 - F(x)]^n \qquad \text{(voir section 5.6)} \\ &= 1 - (1 - x)^n \qquad \text{pour } x \in [0\,,1] \end{aligned}$$

et la fonction de densité est, par dérivation, $f_{X_{(1)}}(x) = n(1-x)^{n-1}$ pour $x \in [0\,,1]$ et 0 sinon. D'où :

$$\begin{aligned} E(X_{(1)}) &= n\int_0^1 x(1-x)^{n-1}dx \quad \text{et avec } x(1-x)^{n-1} = (1-x)^{n-1} - (1-x)^n, \\ &= n\left[\int_0^1 (1-x)^{n-1}dx - \int_0^1 (1-x)^n dx\right] \qquad \text{soit en posant } t = 1-x, \\ &= n\left[\left[\frac{t^n}{n}\right]_0^1 - \left[\frac{t^{n+1}}{n+1}\right]_0^1\right] = n(\frac{1}{n} - \frac{1}{n+1}) = \frac{1}{n+1}. \end{aligned}$$

Exercice 5.13

Reprenons le corrigé de l'exercice 3.7 qui donne $E(X) = 2p-1$, $V(X) = 4p(1-p)$, $E(Y) = n(2p-1)$, $V(Y) = 4np(1-p)$ où $Y = \sum_{i=1}^n X_i$. Pour n grand, Y suit approximativement une loi normale $\mathcal{N}(n(2p-1)\,;4np(1-p))$, donc :

$$\begin{aligned} P(Y > |x|) &= P(-x < Y < x) \\ &\simeq P\left(\frac{-x-n(2p-1)}{2\sqrt{np(1-p)}} < Z < \frac{x-n(2p-1)}{2\sqrt{np(1-p)}}\right) \quad \text{où } Z \rightsquigarrow \mathcal{N}(0\,;1) \\ &= \Phi\left(\frac{x-n(2p-1)}{2\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-x-n(2p-1)}{2\sqrt{np(1-p)}}\right) \end{aligned}$$

où Φ est la fonction de répartition de la loi normale centrée-réduite donnée en section 4.2.4.

Exercices appliqués

Exercice 5.14

Soit X le niveau de bruit d'une machine prise au hasard et $\overline{X}_{10}$ la moyenne d'un échantillon aléatoire de taille 10. On suppose que l'approximation gaussienne s'applique avec $n = 10$ (ce qui est réaliste s'agissant d'une mesure physique dans un processus de fabrication). On a :

$$\overline{X}_{10} \underset{approx}{\rightsquigarrow} \mathcal{N}(44\,;\frac{5^2}{10}),$$

d'où $P(\overline{X}_{10} > 48) \simeq P(Z > \frac{48-44}{5/\sqrt{10}}) = P(Z > 2,53) = 1 - 0,9943 = 0,0057$, où $Z \rightsquigarrow \mathcal{N}(0\,;1)$. On note que cette probabilité est très faible alors que pour une seule machine elle serait $P(Z > \frac{48-44}{5})$ soit environ $0,21$.

Exercice 5.15

$$t = \frac{6,42 - 6,30}{0,22/\sqrt{30}} = 2,99.$$

Pour T, v.a. de Student à 29 degrés de liberté, $P(T < 2,99) = 0,9972$, valeur obtenue dans EXCEL par $1 - LOI.STUDENT(2,99\,;29\,;1)$. On peut vérifier grossièrement l'ordre de grandeur dans la table fournie dans cet ouvrage. La valeur observée correspond au quantile 0,997 ce qui est extrême sur la distribution des valeurs observables et rend l'indication du constructeur peu plausible.

Exercice 5.16

Soit X le poids d'une personne prise au hasard et $\overline{X}_{100}$ la moyenne d'un échantillon aléatoire de taille 100. L'approximation gaussienne s'applique sans problème. On a :

$$\overline{X}_{100} \underset{approx}{\rightsquigarrow} \mathcal{N}\left(66,3\,; \frac{(15,6)^2}{100}\right),$$

d'où $P(\overline{X}_{100} > \frac{7\,000}{100}) \simeq P(Z > \frac{70-66,3}{15,6/\sqrt{100}}) = P(Z > 2,37) = 1 - 0,9911 = 0,0089$, où $Z \rightsquigarrow \mathcal{N}(0\,;1)$.

Exercice 5.17

On suppose que les 1000 personnes sont choisies au hasard dans la population française. Le taux de sondage étant très faible (voir section 3.7), on peut assimiler ce sondage à un sondage avec remise. À chaque tirage la probabilité d'obtenir une personne favorable est 0,44 et, donc, on a :

$$S_n \rightsquigarrow \mathcal{B}(1\,000\,;0,44)$$

d'où $P(S_n \leq 420) \simeq P(U \leq 420,5)$ où $U \rightsquigarrow \mathcal{N}(440\,;246,4)$.
Soit $P(S_n \leq 420) \simeq P(Z \leq \frac{420,5-440}{\sqrt{246,4}}) = P(Z \leq -1,24) = 1 - 0,8925 \simeq 0,11$ où $Z \rightsquigarrow \mathcal{N}(0\,;1)$.

Exercice 5.18

Soit S_n le nombre de pièces défectueuses parmi 120 pièces sélectionnées. On a :

$$S_n \rightsquigarrow \mathcal{B}(120\,;\,0,09)$$

d'où $P(S_n \leq 22) \simeq P(U \leq 22,5)$ où $U \rightsquigarrow \mathcal{N}(10,8\,;\,9,828)$. Soit $P(S_n \leq 22) \simeq P(Z \leq \frac{22,5-10,8}{\sqrt{9,828}}) = P(Z \leq 3,73)$, où $Z \rightsquigarrow \mathcal{N}(0\,;1)$. Cette probabilité est supérieure à 0,9995 comme on peut le voir dans la table (plus exactement elle est égale à 0,99990 selon EXCEL). Cette valeur est très extrême sur la distribution théorique et incite à conclure avec quasi certitude que le fonctionnement est anormal.

Exercice 5.19

Cet exercice, comme d'ailleurs d'autres ci-dessus, préfigure la démarche d'un test statistique. En arrière-plan de l'énoncé, on peut supposer que l'on a observé un écart-type de 0,077 sur cinq mesures faites sur le même échantillon de sang − soit une valeur de précision correspondante de 0,154 mg/l − et l'on souhaite savoir si cette valeur observée est plausible au regard de la distribution théorique de S considérant ce que l'on sait de la méthode de mesure.

Soit donc S l'écart-type d'un échantillon gaussien de taille 5. On s'intéresse à l'événement $(S > 0,077)$ qui équivaut à $(S^2 > 0,005929)$ ou $(\frac{4S^2}{(0,05)^2} > 9,49)$. La statistique $\frac{4S^2}{(0,05)^2}$ correspond ici à $\frac{(n-1)S^2}{\sigma^2}$ qui, selon le théorème 5.1, suit une loi $\chi^2(4)$. On lit dans EXCEL (LOI.KHIDEUX(9,49 ;4)) une probabilité de 0,05 ce qui n'est pas très plausible, sans plus (quantile 0,95). Notons au passage qu'EXCEL n'a pas de cohérence pour les probabilités restituées par les lois et qu'il y a lieu de vérifier avec un exemple de calcul à quoi correspond la probabilité donnée pour telle loi en fonction de x (plus grand que, plus petit que, autre ?).

Exercice 5.20

On suppose que les 10 000 ménages ont des comportements indépendants ce qui donne lieu à une suite de variables aléatoires i.i.d. de loi $\mathcal{P}(4)$. Leur total

T suit une loi $\mathcal{P}(40\,000)$ dont la moyenne est $E(T) = 40\,000$ et $V(T) = 40\,000$. Cette loi peut être précisément approchée par une loi $\mathcal{N}(40\,000\,;\,40\,000)$.

Comme $P(-1,96 < Z < 1,96) = 0,95$ où $Z \rightsquigarrow \mathcal{N}(0\,;1)$, de même $P(40\,000 - 1,96\sqrt{40\,000} < U < 40\,000 + 1,96\sqrt{40\,000}) = 0,95$ où $U \rightsquigarrow \mathcal{N}(40\,000\,;\,40\,000)$, soit $P(39\,608 < U < 40\,392) = 0,95$. Donc, en négligeant toute correction de continuité et en arrondissant, un intervalle de probabilité 0,95 de $[39\,600,\ 40\,400]$.

Exercice 5.21

Pour une seule opération l'erreur d'arrondi est une variable aléatoire $X \rightsquigarrow \mathcal{U}[-\frac{1}{2}, +\frac{1}{2}]$ d'où $E(X) = 0$ et $V(X) = \frac{1}{12}$. Pour T, l'erreur d'arrondi totale sur 1 000 opérations, on a $E(T) = 0$, $V(T) = \frac{1\,000}{12}$ et $T \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;\,\frac{1\,000}{12})$. Comme $P(-1,96 < Z < 1,96) = 0,95$ où $Z \rightsquigarrow \mathcal{N}(0\,;1)$, de même :

$$P(0 - 1,96\sqrt{\frac{1\,000}{12}} < T < 0 + 1,96\sqrt{\frac{1\,000}{12}}) \simeq 0,95$$

soit approximativement un intervalle $[-18, +18]$ en centimes d'euros.

Exercice 5.22

Rappelons que le paramètre λ de la loi $\mathcal{E}(\lambda)$ est l'inverse de sa moyenne. Pour un accumulateur quelconque, on a ainsi une durée de vie $X \rightsquigarrow \mathcal{E}(\frac{1}{2})$ dont la fonction de répartition en x est $F_X(x) = 1 - e^{-\frac{1}{2}x}$ pour $x \geq 0$. L'appareil fonctionne si et seulement si, avec des notations évidentes, l'événement $(X_{(1)} \geq x)$ est réalisé, où $X_{(1)} = \min\{X_1, X_2, X_3\}$. Selon la loi générale du minimum d'un échantillon explicitée en section 5.6, la fonction de répartition de $X_{(1)}$ est, au point $x \geq 0$, $F_{X_{(1)}}(x) = 1 - (e^{-\frac{1}{2}x})^3 = 1 - e^{-\frac{3}{2}x}$ laquelle correspond à la loi $\mathcal{E}(\frac{3}{2})$.

Donc $E(X_{(1)}) = \frac{2}{3}$ et $P(X_{(1)} > 1) = e^{-\frac{3}{2}} \simeq 0,22$.

Exercice 5.23

On doit envisager un processus de Bernoulli avec probabilité de succès 0,9, la variable d'intérêt Y étant le nombre d'essais pour arriver à 100 succès.

La loi de Y est la loi binomiale négative sous la deuxième forme définie en fin de section 4.1.4, avec paramètres $r = 100$ et $p = 0,9$. Alors Y prend ses valeurs dans $\{100, 101, \cdots\}$, $E(Y) = \frac{r}{p} \simeq 111,1$ et $V(Y) = \frac{r(1-p)}{p^2} \simeq 12,35$. Comme on a vu dans la dite section qu'une loi $\mathcal{BN}(r,p)$ peut être envisagée comme une somme de r variables aléatoires i.i.d. de loi $\mathcal{G}(p)$, le théorème central limite s'applique et la loi binomiale négative peut être approchée par une loi de Gauss pourvu que r soit suffisamment grand, ce qui est le cas ici car r est largement supérieur à 30. Ainsi, en utilisant la correction de continuité :

$$\begin{aligned} P(Y > 111) &\simeq P(U > 111,5) \text{ où } U \rightsquigarrow \mathcal{N}(111,1\,;\,12,35) \\ &\simeq P(Z > \frac{111,5 - 111,1}{\sqrt{12,35}}) = P(Z > 0,11) = 1 - 0,5438 \simeq 0,46. \end{aligned}$$

Pour être sûr de fabriquer 100 pièces bonnes, il faudrait prévoir de fabriquer une infinité de pièces !

Pour atteindre 100 bonnes pièces avec probabilité $0,99$ il faut fabriquer n pièces où n est tel que $P(Y \leq n) = 0,99$ ou plus correctement $P(Y \leq n) \geq 0,99$ puisque nous sommes sur une loi discrète. Or $P(Z < 2,33) = 0,99$, donc $P(U < 111,1 + 2,33\sqrt{12,35}) = 0,99$, soit $P(U < 119,29) = 0,99$. Comme $119,29 < 119,5$ on prendra $n = 120$ pièces.

Chapitre 6 : Théorie de l'estimation paramétrique ponctuelle

Exercice 6.1

Il s'agit de voir si ces lois à un paramètre (inconnu), disons θ, ont une fonction de probabilité, respectivement fonction de densité de probabilité, pouvant se mettre sous la forme (voir section 6.3) :

$$p(x;\theta) = a(\theta)b(x)\exp\{c(\theta)d(x)\}$$

où a, b, c, d sont des fonctions.

Loi $BN(r,p)$ avec r connu

Fonction de probabilité :

$$\begin{aligned} p(x;p) &= \binom{r+x-1}{x} p^r(1-p)^x,\ x \in \mathbb{N} \\ &= \binom{r+x-1}{x} p^r \exp\{x\ln(1-p)\} \end{aligned}$$

forme classe exponentielle avec $a(p) = p^r$, $b(x) = \binom{r+x-1}{x}$, $c(p) = \ln(1-p)$ et $d(x) = x$.

Loi $P(\lambda)$

Fonction de probabilité :

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left(\frac{1}{x!}\right)\exp\{x\ln\lambda\}$$

forme classe exponentielle avec $a(\lambda) = e^{-\lambda}$, $b(x) = \frac{1}{x!}$, $c(\lambda) = \ln\lambda$ et $d(x) = x$.

Loi $E(\lambda)$

Fonction de densité de probabilité :

$$f(x;\lambda) = \lambda e^{-\lambda x} = \lambda\exp\{-\lambda x\}$$

forme classe exponentielle avec $a(\lambda) = \lambda$, $b(x) = 1$, $c(\lambda) = -\lambda$ et $d(x) = x$.

Loi $\Gamma(r,\lambda)$ avec r connu

Fonction de densité de probabilité :

$$f(x;\lambda)=\frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}=\left(\frac{\lambda^r}{\Gamma(r)}\right)(x^{r-1})\exp\{-\lambda x\}$$

forme classe exponentielle avec $a(\lambda)=\left(\frac{\lambda^r}{\Gamma(r)}\right)$, $b(x)=x^{r-1}$, $c(\lambda)=-\lambda$ et $d(x)=x$.

Exercice 6.2

$$f(x;\theta)=\theta a^\theta x^{-(\theta+1)}I_{[a,+\infty[}(x)=(\theta a^\theta)I_{[a,+\infty[}(x)\exp\{-(\theta+1)\ln x\},$$

de forme classe exponentielle avec $a(\theta)=\theta a^\theta$, $b(x)=I_{[a,+\infty[}(x)$, $c(\theta)=-(\theta+1)$ et $d(x)=\ln x$.

Donc $\sum_{i=1}^n \ln X_i$ est statistique exhaustive minimale.

Comme $\mu=\frac{\theta a}{\theta-1}$ (voir section 4.2.6), l'estimateur des moments $\widehat{\theta}^M$est tel que $\overline{X}=\frac{\widehat{\theta}^M a}{\widehat{\theta}^M-1}$, soit $\widehat{\theta}^M=\frac{\overline{X}}{\overline{X}-a}$. On constate qu'il n'est pas fonction de la statistique exhaustive minimale ci-dessus (et en ce sens il ne saurait être un des plus pertinents).

Exercice 6.3

$$\begin{aligned}f(x;\alpha,\beta)&=\frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+2)}x^\alpha(1-x)^\beta I_{[0,1]}(x)\\&=\frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+1)\Gamma(\beta+1)}I_{]0,1[}(x)\exp\{\alpha\ln x+\beta\ln(1-x)\},\end{aligned}$$

de forme classe exponentielle avec $d_1(x)=\ln x$ et $d_2(x)=\ln(1-x)$. Donc le couple $(\sum_{i=1}^n \ln X_i, \sum_{i=1}^n \ln(1-X_i))$ est statistique exhaustive minimale.

Pour l'estimateur des moments du couple (α,β), on utilise le fait que (voir section 4.2.8) $E(X)=\frac{\alpha+1}{\alpha+\beta+2}$ et $E(X^2)=\frac{(\alpha+1)(\beta+1)}{(\alpha+\beta+2)^2(\alpha+\beta+3)}+\frac{(\alpha+1)^2}{(\alpha+\beta+2)^2}=\frac{(\alpha+1)(\alpha+2)}{(\alpha+\beta+2)(\alpha+\beta+3)}$. Il s'agit donc de résoudre en (α,β) le système :

$$\left\{\begin{array}{c}\frac{\alpha+1}{\alpha+\beta+2}=\overline{X}\\ \frac{(\alpha+1)(\alpha+2)}{(\alpha+\beta+2)(\alpha+\beta+3)}=\frac{1}{n}\sum_{i=1}^n X_i^2\end{array}\right.\Leftrightarrow\left\{\begin{array}{c}\frac{\alpha+1}{\alpha+\beta+2}=\overline{X}\\ \frac{\alpha+2}{\alpha+\beta+3}\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i^2\end{array}\right.,$$

soit deux équations linéaires en α et β, dont la solution *in fine* est :

$$\widehat{\alpha}^M = \frac{\overline{X}^2(1-\overline{X}) - \widetilde{S}^2(1+\overline{X})}{\widetilde{S}^2} \quad \text{et} \quad \widehat{\beta}^M = \frac{\overline{X} - (2-\overline{X})(\widetilde{S}^2 + \overline{X}^2)}{\widetilde{S}^2}.$$

On constate qu'elle n'est pas fonction de la statistique exhaustive minimale ci-dessus (et en ce sens elle ne saurait être une des plus pertinentes).

Exercice 6.4

La fonction de répartition de S en s est $P(S \leq s) = P(S^2 \leq s^2) = P\left(\frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)s^2}{\sigma^2}\right) = F_{\chi^2(n-1)}\left(\frac{(n-1)s^2}{\sigma^2}\right)$ (selon le théorème 5.1) où $F_{\chi^2(n-1)}$ est la fonction de répartition d'une loi du khi-deux à $(n-1)$ degrés de liberté. La fonction de densité de S est donc : $\frac{2(n-1)s}{\sigma^2} f_{\chi^2(n-1)}\left(\frac{(n-1)s^2}{\sigma^2}\right)$ que l'on peut écrire explicitement en substituant (proposition 5.5) $f_{\chi^2(n-1)}(x) = [2^{(n-1)/2}\Gamma(\frac{n-1}{2})]^{-1} x^{(n-3)/2} e^{-x/2}$, avec $x > 0$. On peut calculer directement $E(S)$ par :

$$E(S) = \frac{\sigma}{\sqrt{n-1}} E\left(\left[\frac{(n-1)S^2}{\sigma^2}\right]^{\frac{1}{2}}\right) = \frac{\sigma}{\sqrt{n-1}} \frac{1}{2^{(n-1)/2}\Gamma(\frac{n-1}{2})} \int_0^{+\infty} x^{\frac{1}{2}} x^{\frac{n-3}{2}} e^{-\frac{x}{2}} dx$$

où l'intégrale n'est autre que la densité de la loi du khi-deux à n degrés de liberté au facteur $[2^{n/2}\Gamma(\frac{n}{2})]^{-1}$ près. D'où :

$$E(S) = \sigma\sqrt{\frac{2}{n-1}}\ \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}.$$

Pour éliminer le biais, il suffit de prendre, comme estimateur de σ, $\sqrt{\frac{n-1}{2}}\frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n}{2})} S$.

Exercice 6.5

1. Soit $n = 2k-1$, la médiane est la statistique d'ordre k, notée $X_{(k)}$. Sa fonction de répartition (voir proposition 5.12) est :
$G(x;\theta) = \sum_{j=k}^{n} \binom{n}{j} [F(x-\theta)]^j [1-F(x-\theta)]^{n-j}$. Sa densité est donc :
$g(x;\theta) = \sum_{j=k}^{n} \binom{n}{j} f(x-\theta)\{\, j[F(x-\theta)]^{j-1}[1-F(x-\theta)]^{n-j}$
$-(n-j)[F(x-\theta)]^j[1-F(x-\theta)]^{n-j-1}\}$
$g(x;\theta) = \sum_{j=k}^{n} \binom{n}{j} f(x-\theta)[F(x-\theta)]^{j-1}[1-F(x-\theta)]^{n-j-1}[j - nF(x-\theta)]$.

2. Notons que $f(-t) = f(t)$ équivaut à $F(-t) = 1 - F(t)$. Soit $x_1 = \theta + h$ et $x_2 = \theta - h$. Il suffit de montrer que $G(x_2;\theta) = 1 - G(x_1;\theta)$. Posons $F(h) = a$.

Ainsi :

$$G(x_1;\theta) = \sum_{j=k}^{n} \binom{n}{j} a^j (1-a)^{n-j}$$

$$G(x_2;\theta) = \sum_{j=k}^{n} \binom{n}{j} [F(-h)]^j [1-F(-h)]^{n-j} = \sum_{j=k}^{n} \binom{n}{j} (1-a)^j a^{n-j}$$

$$= \sum_{s=0}^{n-k} \binom{n}{n-s} (1-a)^{n-s} a^s = \sum_{s=0}^{k-1} \binom{n}{s} a^s (1-a)^{n-s}$$

en posant $s = n - j$ (et comme $n = 2k - 1$).
Donc $G(x_1;\theta) + G(x_2;\theta) = \sum_{t=1}^{n} \binom{n}{t} a^t (1-a)^{n-t} = 1$.

3. Soit M la médiane empirique, alors $E(M-\theta) = \int_{-\infty}^{+\infty} (x-\theta) g(x;\theta) dx = \int_{-\infty}^{+\infty} t g(t+\theta;\theta) dt = 0$ car $g(-t+\theta;\theta) = g(t+\theta;\theta)$. Notons qu'il est toutefois nécesssaire que ces intégrales existent, ce qui revient à ce que $\int_{-\infty}^{+\infty} x f(x-\theta) dx$ existe, soit que la loi mère ait une espérance mathématique.

Exercice 6.6

$\overline{X}$ est sans biais pour $\frac{1}{\lambda}$ puisque $E(\overline{X}) = \frac{1}{\lambda}$; $V(\overline{X}) = \frac{1}{n\lambda^2}$; $eqm_{\frac{1}{\lambda}}(\overline{X}) = \frac{1}{n\lambda^2}$.
$T = \frac{1}{n+1} \sum_{i=1}^{n} X_i = \frac{n}{n+1} \overline{X}$, donc $E(T) = \frac{n}{n+1} \frac{1}{\lambda}$ et T a un biais $-\frac{1}{n+1} \frac{1}{\lambda}$; $V(T) = \frac{n}{(n+1)^2} \frac{1}{\lambda^2}$.
D'où $eqm_{\frac{1}{\lambda}}(T) = \frac{1}{(n+1)^2} \frac{1}{\lambda^2} + \frac{n}{(n+1)^2} \frac{1}{\lambda^2} = \frac{1}{(n+1)^2 \lambda^2} < eqm_{\frac{1}{\lambda}}(\overline{X})$. En erreur quadratique moyenne T est meilleur, le gain de variance étant supérieur à la perte due au biais.

Exercice 6.7

Comme vu à l'exercice 5.4, $V(S^2) = \frac{1}{n}(\mu_4' - \frac{n-3}{n-1}\sigma^4)$ et est sans biais donc son e.q.m. est $V(S^2)$.
$\widetilde{S}^2 = \frac{n-1}{n} S^2$ et a un biais $-\frac{\sigma^2}{n}$; $V(\widetilde{S}^2) = \frac{(n-1)^2}{n^2} V(S^2) = \frac{(n-1)^2}{n^3}(\mu_4' - \frac{n-3}{n-1}\sigma^4)$
$\widetilde{S}^2$ domine S^2 en e.q.m. si $\frac{1}{n}(\mu_4' - \frac{n-3}{n-1}\sigma^4) - \left[\frac{(n-1)^2}{n^3}(\mu_4' - \frac{n-3}{n-1}\sigma^4) + \frac{\sigma 4}{n^2}\right] > 0$, soit $\frac{2n-1}{n}(\mu_4' - \frac{n-3}{n-1}\sigma^4) > \frac{\sigma 4}{n^2}$ ou, finalement, $\mu_4' > \frac{3n^2-8n+3}{(2n-1)(n-1)}\sigma 4$.
Pour la loi de Gauss $\mu_4' = 3\sigma 4$ et l'on peut aisément vérifier que cette condition est remplie pour tout $n > 1$. Si l'on s'en tient au critère de l'e.q.m., on doit préférer $\widetilde{S}^2$.

Exercice 6.8

Soit pour $X \rightsquigarrow \mathcal{U}[0,\theta]$ la fonction de répartition $F_X(x;\theta) = \frac{x}{\theta}$ si $x \in [0,\theta]$. Donc pour un échantillon de taille n, $F_{X_{(n)}}(x;\theta) = [F_X(x;\theta)]^n = \left(\frac{x}{\theta}\right)^n$ si $x \in [0,\theta]$. Pour tout $\varepsilon > 0$, $P(\left|X_{(n)} - \theta\right| > \varepsilon) = P(X_{(n)} < \theta - \varepsilon) = \left(\frac{\theta-\varepsilon}{\theta}\right)^n$ qui tend vers 0 quand $n \to \infty$. Ainsi pour tout $\varepsilon > 0$, $P(\left|X_{(n)} - \theta\right| < \varepsilon) \to 1$ quand $n \to \infty$ ce qui définit la convergence en probabilité de $X_{(n)}$ vers θ (définition 5.14).

Exercice 6.9

L'estimateur UMVUE est $T = \frac{n-1}{\sum_{i=1}^n X_i}$ (voir exemple 6.11). Posons $T_n = \sum_{i=1}^n X_i$. Suivant le même argument que pour démontrer $E(\frac{1}{T_n}) = \frac{\lambda}{n-1}$ dans l'exemple 6.11, on a, à condition que $n > 2$:

$$E(\frac{1}{T_n^2}) = \int_0^{+\infty} \frac{1}{t^2} \frac{\lambda}{(n-1)!}(\lambda t)^{n-1} e^{-\lambda t} dt$$
$$= \frac{\lambda^2}{(n-1)(n-2)} \int_0^{+\infty} \frac{\lambda}{(n-3)!}(\lambda t)^{n-3} e^{-\lambda t} dt = \frac{\lambda^2}{(n-1)(n-2)}.$$

Donc $V(\frac{1}{T_n}) = \frac{\lambda^2}{(n-1)(n-2)} - \frac{\lambda^2}{(n-1)^2} = \frac{\lambda^2}{(n-1)^2(n-2)}$, d'où $V(T) = \frac{\lambda^2}{(n-2)} = eqm_\lambda(T)$ pour l'estimateur UMVUE.

L'estimateur des moments est $\frac{1}{\overline{X}} = \frac{n}{n-1}T$, d 'où $E(\frac{1}{\overline{X}}) = \frac{n}{n-1}\lambda$ avec un biais $\frac{\lambda}{n-1}$; $V(\frac{1}{\overline{X}}) = \left(\frac{n}{n-1}\right)^2 \frac{\lambda^2}{(n-2)}$ et :

$$eqm_\lambda(\frac{1}{\overline{X}}) = \frac{1}{(n-1)^2}\lambda^2 + \frac{n^2}{(n-1)^2(n-2)}\lambda^2 = \frac{n^2+n-2}{(n-1)^2(n-2)}\lambda^2.$$

Or, pour tout $n > 2$, $\frac{n^2+n-2}{(n-1)^2} > 1$ car $n^2 + n - 2 - (n-1)^2 = 3(n-1)$. Non seulement l'estimateur des moments est dominé en e.q.m., mais en plus il est biaisé.

Exercice 6.10

De toute évidence, la famille est dans la classe exponentielle, avec $d(x) = x^2$. Ainsi $\sum_{i=1}^n d(X_i) = \sum_{i=1}^n X_i^2$ est statistique exhaustive minimale.
$E(X^2) = \int_0^{+\infty} \frac{x^3}{\theta} e^{-\frac{x^2}{2\theta}} dx$, soit, en posant $t = \frac{x^2}{2\theta}$, $E(X^2) = 2\theta \int_0^{+\infty} te^{-t} dt$. Or l'intégrale ci-dessus est la moyenne de la loi $\mathcal{E}(1)$ et vaut donc 1. Donc

$E(X^2) = 2\theta$ et $E(\frac{1}{2n}\sum_{i=1}^n X_i^2) = \theta$. En vertu de la proposition 6.7, $\frac{1}{2n}\sum_{i=1}^n X_i^2$ − sans biais et fonction linéaire de $\sum_{i=1}^n d(X_i)$ − est efficace. On peut dire aussi que cet estimateur de θ est UMVUE.

Exercice 6.11

1. $P(\ln\left(\frac{X}{a}\right) < x) = P(X < ae^x) = 1 - \left(\frac{a}{ae^x}\right)^\theta$ si $ae^x \geq a$, soit $x \geq 0$ (voir fonction de répartition de la loi de Pareto en section 4.2.6). Ainsi $P(\ln\left(\frac{X}{a}\right) < x) = 1 - e^{-\theta x}$ pour $x \geq 0$, qui est la fonction de répartition de la loi $\mathcal{E}(\theta)$.

2. La densité de la loi de Pareto est :
$f(x;\theta) = \theta a^\theta x^{-(\theta+1)} I_{[a,+\infty[}(x) = \theta a^\theta I_{[a,+\infty[}(x) \exp\{-(\theta+1)\ln x\}$,
qui met en évidence la forme de la classe exponentielle avec $d(x) = \ln x$.
Posons $Y_i = \ln(\frac{X_i}{a})$, alors $\sum_{i=1}^n Y_i \rightsquigarrow \Gamma(n,\theta)$ puisque $Y_i \rightsquigarrow \mathcal{E}(\theta)$.
Selon l'exemple 6.11, $E\left(\frac{1}{\sum_{i=1}^n Y_i}\right) = \frac{\theta}{n-1}$, d'où $E\left(\frac{n-1}{\sum_{i=1}^n Y_i}\right) = \theta$. La statistique

$$\frac{n-1}{\sum_{i=1}^n \ln(\frac{X_i}{a})} = \frac{n-1}{\sum_{i=1}^n \ln X_i - n\ln a}$$

est une fonction de $\sum_{i=1}^n \ln X_i$ sans biais pour θ. D'après la proposition 6.6 c'est la statistique UMVUE pour estimer θ.

3. Selon la proposition 6.7, seul $\sum_{i=1}^n d(X_i)$ est efficace pour estimer $h(\theta) = E_\theta(\sum_{i=1}^n d(X_i))$ et ceci à une fonction linéaire près. Or $E(\ln(\frac{X}{a})) = \frac{1}{\theta}$, la moyenne de la loi $\mathcal{E}(\theta)$. Ainsi $E(\frac{1}{n}\sum_{i=1}^n \ln(\frac{X_i}{a})) = \frac{1}{\theta}$ et $\frac{1}{n}\sum_{i=1}^n \ln(\frac{X_i}{a})$ est sans biais et efficace pour estimer $\frac{1}{\theta}$.

Exercice 6.12

1. La densité est :
$f(x;\theta) = \theta a^\theta x^{-(\theta+1)} I_{[a,+\infty[}(x)$ qui ne peut être mis sous la forme de la classe exponentielle en raison de $I_{[a,+\infty[}(x)$.

2. L'estimateur des moments pour a est solution de $\frac{a\theta}{\theta-1} = \overline{X}$ (voir section 4.2.6, si $\theta > 1$), soit $\widehat{a}^M = \frac{\theta-1}{\theta}\overline{X}$.

3. Utilisons le théorème 6.1 :

$$\prod_{i=1}^n f(x_i;a) = \theta^n a^{n\theta}\left(\prod_{i=1}^n x_i\right)^{-(\theta+1)} \prod_{i=1}^n I_{[a,+\infty[}(x_i),$$

or $\prod_{i=1}^{n} I_{[a,+\infty[}(x_i) = \prod_{i=1}^{n} I_{[a,+\infty[}(x_{(1)})$ où $x_{(1)} = \min\{x_1, \cdots, x_n\}$. Donc :

$$\prod_{i=1}^{n} f(x_i; a) = a^{n\theta} \prod_{i=1}^{n} I_{[a,+\infty[}(x_{(1)}) \left(\prod_{i=1}^{n} x_i\right)^{-(\theta+1)} \theta^n$$

et d'après le théorème $X_{(1)}$ est exhaustive. Elle est minimale du fait qu'elle est de dimension 1.

4. La fonction de répartition de $X_{(1)}$ est $1-[1-F(x;a)]^n = 1-\left(\frac{a}{x}\right)^{n\theta}$ si $x \geq a$. Donc $X_{(1)}$ suit une loi de Pareto de paramètre de seuil a et de paramètre de forme $n\theta$, d'où $E\left(X_{(1)}\right) = \frac{n\theta}{n\theta-1}a$, $E\left(\frac{n\theta-1}{n\theta}X_{(1)}\right) = a$ et $(1-\frac{1}{n\theta})X_{(1)}$ est un estimateur sans biais pour a. Notons qu'il n'est pas pour autant UMVUE car on n'est pas dans la classe exponentielle.

Exercice 6.13

1. Pour la famille des lois $\mathcal{N}(\mu\,, 1)$, $\mu \in \mathbb{R}$, on a la densité :
$f(x;\mu) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}\,(x-\mu)^2\}$. Calculons :
$\ln f(x;\mu) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\,(x-\mu)^2$, $\frac{\vartheta}{\vartheta\mu}\ln f(x;\mu) = x-\mu$,
$\frac{\vartheta^2}{\vartheta\mu^2}\ln f(x;\mu) = -1$, $I(\mu) = E\left[-\frac{\vartheta^2}{\vartheta\mu^2}\ln f(X;\mu)\right] = 1$.
La borne de Cramer-Rao est $\frac{1}{nI(\mu)} = \frac{1}{n}$. Comme $\overline{X}$ est sans biais pour μ et que $V(\overline{X}) = \frac{1}{n}$, $\overline{X}$ est efficace.

2. Pour la famille des lois $\mathcal{N}(0\,, \sigma^2)$, $\sigma^2 > 0$, posons $v = \sigma^2$ pour simplifier les écritures.
$f(x;v) = \frac{1}{\sqrt{2\pi v}} \exp\{-\frac{1}{2}\frac{x^2}{v}\}$, $\ln f(x;v) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln v - \frac{1}{2}\frac{x^2}{v}$,
$\frac{\vartheta}{\vartheta v}\ln f(x;v) = -\frac{1}{2v} + \frac{x^2}{2v^2}$, $\frac{\vartheta^2}{\vartheta v^2}\ln f(x;v) = \frac{1}{2v^2} - \frac{x^2}{v^3}$,
$I(v) = E\left[-\frac{\vartheta^2}{\vartheta v^2}\ln f(X;v)\right] = -\frac{1}{2v^2} + \frac{E(X^2)}{v^3} = -\frac{1}{2v^2} + \frac{1}{v^2}$,
car $E(X^2) = \sigma^2 = v$, donc $I(v) = \frac{1}{2v^2}$ ou $I(\sigma^2) = \frac{1}{2\sigma^4}$ et la borne de Cramer-Rao est $\frac{1}{nI(\sigma^2)} = \frac{2\sigma^4}{n}$.
Notons que $\frac{\sum_{i=1}^{n} X_i^2}{n}$ est sans biais pour σ^2et que $V\left(\frac{\sum_{i=1}^{n} X_i^2}{n}\right) = \frac{1}{n}V(X^2) = \frac{1}{n}\left[E\left(X^4\right) - \left[E\left(X^2\right)\right]^2\right] = \frac{1}{n}\left(3\sigma^4 - \sigma^4\right) = \frac{2\sigma^4}{n}$. Donc cet estimateur est efficace.

Exercice 6.14

Comme vu en section 6.6.3, pour estimer une fonction $h(p) = \frac{p}{1-p}$ du paramètre p, la borne de Cramer-Rao est $\frac{[h'(p)]^2}{nI(p)} = \frac{1}{nI(p)(1-p)^4}$. Pour la loi de Bernoulli $f(x;p) = p^x(1-p)^{1-x}$, $\ln f(x;p) = x\ln p + (1-x)\ln(1-p)$,
$\frac{\vartheta}{\vartheta p}\ln f(x;p) = \frac{x}{p} - \frac{1-x}{1-p}$, $\frac{\vartheta^2}{\vartheta p^2}\ln f(x;p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}$,
$I(p) = E\left[-\frac{\vartheta^2}{\vartheta p^2}\ln f(X;p)\right] = \frac{E(X)}{p^2} + \frac{1-E(X)}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$. La borne de Cramer-Rao est donc, pour estimer $h(p)$, $\frac{p}{n(1-p)^3}$.

Exercice 6.15

On a : $\ln f(x;\theta) = -\ln\pi - \ln[1+(x-\theta)^2]$, $\frac{\vartheta}{\vartheta\theta}\ln f(x;\theta) = \frac{-2(x-\theta)}{1+(x-\theta)^2}$,
$I(\theta) = E\left[\left(\frac{\vartheta}{\vartheta\theta}\ln f(X;\theta)\right)^2\right] = E\left[\frac{4(X-\theta)^2}{[1+(X-\theta)^2]^2}\right] = \frac{4}{\pi}\int_{-\infty}^{+\infty}\frac{(x-\theta)^2}{[1+(x-\theta)^2]^3}dx$.
Posons $t = x - \theta$:
$I(\theta) = \frac{4}{\pi}\int_{-\infty}^{+\infty}\frac{t^2}{(1+t^2)^3}dt = \frac{4}{8\pi}\left[\arctan t + \frac{t(t^2-1)}{(t^2+1)^2}\right]_{-\infty}^{+\infty} = \frac{1}{2\pi}\pi = \frac{1}{2}$.
La borne de Cramer-Rao pour θ est donc $\frac{2}{n}$.

La variance asymptotique de $\sqrt{n}(M_n - \theta)$ est $\frac{\pi^2}{4} \simeq 2,47 > 2$, donc M_n n'est pas asymptotiquement efficace et n'est pas un estimateur BAN (voir la définition en proposition 6.11).

Exercice 6.16

Estimateur des moments

Il vérifie $\frac{\theta}{2} = \overline{X}$ et c'est donc $\widehat{\theta}^M = 2\overline{X}$. Il est sans biais et $eqm(\widehat{\theta}^M) = V(\widehat{\theta}^M) = 4V(\overline{X}) = \frac{4}{n}V(X) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}$.

Estimateur du MV

On a vu que $\widehat{\theta}^{MV} = X_{(n)}$ (exemple 6.21) et que $E(X_{(n)}) = \frac{n}{n+1}\theta$ (exemple 6.4). Son biais est $\frac{n}{n+1}\theta - \theta = -\frac{\theta}{n+1}$. Calculons $V(X_{(n)}) = E(X_{(n)}^2) - \frac{n^2}{(n+1)^2}\theta^2$. La densité de $X_{(n)}$ étant $n\frac{x^{n-1}}{\theta^n}$ si $x \in [0,\theta]$ et 0 sinon (voir exemple 6.4), on a :
$E(X_{(n)}^2) = n\int_0^\theta \frac{x^{n+1}}{\theta^n}dx = \frac{n}{n+2}\theta^2$; $V(X_{(n)}) = \frac{n}{(n+1)^2(n+2)}\theta^2$. D'où :
$eqm(X_{(n)}) = \left(-\frac{\theta}{n+1}\right)^2 + \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{2}{(n+1)(n+2)}\theta^2$.

Estimateur $\frac{n+1}{n}X_{(n)}$ (UMVUE)

Calculons sa variance d'après celle de l'EMV.

$eqm(\frac{n+1}{n}X_{(n)}) = V(\frac{n+1}{n}X_{(n)}) = \left(\frac{n+1}{n}\right)^2 V(X_{(n)}) = \frac{1}{n(n+2)}\theta^2$.

Pour $n \geq 3$, on vérifie aisément que $eqm(\frac{n+1}{n}X_{(n)}) < eqm(X_{(n)}) < eqm(\widehat{\theta}^M)$.
Pour $n = 2$, la deuxième inégalité devient une égalité.
Notons que pour n grand, $eqm(\frac{n+1}{n}X_{(n)}) \sim \frac{\theta^2}{n^2}$, $eqm(X_{(n)}) \sim \frac{2\theta^2}{n^2}$, $eqm(\widehat{\theta}^M) \sim \frac{\theta^2}{3n}$. L'estimateur des moments est largement dominé par les deux autres (il est peu pertinent car il n'est pas une fonction de $X_{(n)}$, statistique exhaustive minimale).

Exercice 6.17

La fonction de probabilité de cette loi au point k s'obtient par :
$f(x;\lambda) = P(X = k|\ X \neq 0) = \frac{P(X=k)}{1-P(X=0)} = \frac{1}{1-e^{-\lambda}}\frac{e^{-\lambda}\lambda^x}{x!}$. Alors :
$\ln f(x;\lambda) = -\lambda + x\ln\lambda - \ln x! - \ln(1-e^{-\lambda})$.
$\frac{\vartheta}{\vartheta\lambda}\ln f(x;\lambda) = -1 + \frac{x}{\lambda} - \frac{e^{-\lambda}}{1-e^{-\lambda}}$, $\frac{\vartheta}{\vartheta\lambda}\ln L(\lambda) = -n + \frac{\sum_{i=1}^n x_i}{\lambda} - \frac{ne^{-\lambda}}{1-e^{-\lambda}}$.

L'estimation du MV est la valeur de λ (si unique) telle que :
$-n + \frac{\sum_{i=1}^n x_i}{\lambda} - \frac{ne^{-\lambda}}{1-e^{-\lambda}} = 0$ ou $\frac{\lambda}{1-e^{-\lambda}} = \overline{x}$.
Pour $x = 3$, la solution donne, par approximations successives, $\widehat{\lambda}^{MV} \simeq 2,82$.

Exercice 6.18

La densité de la loi de Pareto est :
$f(x;\theta) = \frac{\theta}{a}\left(\frac{a}{x}\right)^{\theta+1}$ pour $x \geq a$ et 0 sinon.
$\ln f(x;\theta) = \ln\frac{\theta}{a} + (\theta+1)\ln\frac{a}{x}$; $\frac{\vartheta}{\vartheta\theta}\ln f(x;\theta) = \frac{1}{\theta} + \ln\frac{a}{x}$;
$\frac{\vartheta}{\vartheta\theta}\ln L(\theta) = \frac{n}{\theta} + \sum_{i=1}^n \ln\frac{a}{x_i}$.

L'estimation du MV est telle que :
$\frac{n}{\theta} + \sum_{i=1}^n \ln\frac{a}{x_i} = 0$, soit pour l'estimateur $\widehat{\theta}^{MV} = \frac{n}{\sum_{i=1}^n \ln\frac{a}{X_i}}$.
Dans la solution de l'exercice 6.11 on a vu que $E\left(\frac{n-1}{\sum_{i=1}^n \ln\frac{a}{X_i}}\right) = \theta$, donc $E(\widehat{\theta}^{MV}) = \frac{n}{n-1}\theta$. Son biais est $\frac{n}{n-1}\theta - \theta = \frac{1}{n-1}\theta$.

Pour la borne de Cramer-Rao calculons : $\frac{\vartheta^2}{\vartheta\theta^2}\ln f(x;\theta) = -\frac{1}{\theta^2}$, d'où $I(\theta) = \frac{1}{\theta^2}$ et la borne est $\frac{\theta^2}{n}$.

Pour la variance de l'EMV notons que selon l'exercice 6.11, $Y_i = \ln(\frac{X_i}{a})$ et $Y_i \rightsquigarrow \mathcal{E}(\theta)$. Posons $T = \sum_{i=1}^n Y_i$, alors $V(\widehat{\theta}^{MV}) = V(\frac{n}{T}) = \frac{n^2}{(n-2)(n-1)^2}\theta^2$

comme il a été établi lors de la solution de l'exercice 6.9 pour $V(\frac{1}{T})$. Comme $\frac{n^2}{(n-2)(n-1)^2}\theta^2 \sim \frac{\theta^2}{n}$ quand $n \to \infty$, $\widehat{\theta}^{MV}$ est asymptotiquement efficace.

Exercice 6.19

Calculons :

$\ln f(x;\rho) = \ln\rho + \ln(\rho+1) + \ln x + (\rho-1)\ln(1-x)$,

$\frac{\vartheta}{\vartheta\rho}\ln f(x;\rho) = \frac{1}{\rho} + \frac{1}{\rho+1} + \ln(1-x)$.

Donc $\widehat{\rho}^{MV}$ (estimation du MV) est solution de $\frac{n}{\rho} + \frac{n}{\rho+1} + \sum_{i=1}^{n}\ln(1-x_i) = 0$ ou $\frac{2\rho+1}{\rho(\rho+1)} = -\frac{1}{n}\sum_{i=1}^{n}\ln(1-x_i)$. Posons $a = -\frac{1}{n}\sum_{i=1}^{n}\ln(1-x_i)$. Il faut résoudre en ρ l'équation $2\rho+1 = a\rho(\rho+1)$. Les solutions sont $\frac{2-a\pm\sqrt{4+a^2}}{2a}$. Comme a ne prend que des valeurs positives, la seule solution dans $\mathbb{R}^+$ est $\widehat{\rho}^{MV} = \frac{2-a+\sqrt{4+a^2}}{2a}$.

Exercice 6.20

Loi binomiale négative $BN(r,p)$ avec r connu

fonction de probabilité $f(x;p) = \binom{r+x-1}{x}p^r(1-p)^x$, $x \in \mathbb{N}$.

$\ln f(x;p) = \ln\binom{r+x-1}{x} + r\ln p + x\ln(1-p)$

$\frac{\vartheta}{\vartheta p}\ln f(x;p) = \frac{r}{p} - \frac{x}{1-p}$; $\widehat{p}^{MV}$ (estimation) solution de $\frac{r}{p} - \frac{x}{1-p} = 0$, d'où $\widehat{p}^{MV} = \frac{r}{X+r}$ (estimateur). La solution est intuitivement réaliste puisqu'elle consiste à faire le rapport du nombre de succès sur le nombre d'essais.

Loi binomiale $B(n,p)$ avec n connu

fonction de probabilité $f(x;p) = \binom{n}{x}p^x(1-p)^{n-x}$, $x = 0, 1, \cdots, n$.

$\ln f(x;p) = \ln\binom{n}{x} + x\ln p + (n-x)\ln(1-p)$

$\frac{\vartheta}{\vartheta p}\ln f(x;p) = \frac{x}{p} - \frac{n-x}{1-p}$; $\widehat{p}^{MV}$ (estimation) solution de $\frac{x}{p} - \frac{n-x}{1-p} = 0$, d'où $\widehat{p}^{MV} = \frac{X}{n}$ (estimateur), nombre de succès sur nombre d'essais.

Exercice 6.21

La fonction de probabilité non nulle pour $x \in \{1, 2, \cdots, N-M+1\}$ est :

$$f(x;N) = P(X=x) = \frac{N-M}{N}\frac{N-M-1}{N-1}\cdots\frac{N-M-(x-2)}{N-(x-2)}\frac{M}{N-(x-1)}.$$

$\ln f(x;N) = \sum_{k=0}^{x-2}\ln(N-M-k) - \sum_{k=0}^{x-1}\ln(N-k) + \ln M$

$\frac{\vartheta}{\vartheta N}\ln f(x;N) = \sum_{k=0}^{x-2}\frac{1}{N-M-k} - \sum_{k=0}^{x-1}\frac{1}{N-k}$; $\widehat{N}^{MV}$ (estimation) solution annulant cette expression.

Application

$M = 100$, $x = 3$, résoudre en N : $\frac{1}{N-100} + \frac{1}{N-101} - (\frac{1}{N} + \frac{1}{N-1} + \frac{1}{N-2}) = 0$. La solution donnée par un logiciel mathématique est $\widehat{N}^{MV} = 300$.

On notera que c'est la même solution que dans un échantillonnage sans remise où $Y = X - 1 \rightsquigarrow \mathcal{G}(p)$ avec $p = \frac{M}{N}$, car $\widehat{p}^{MV} = \frac{1}{y+1} = \frac{1}{x} = \frac{1}{3}$.

Exercice 6.22

La fonction de probabilité est $f(x;\lambda) = \frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}$, $x > 0$.

$\ln f(x;\lambda) = r\ln\lambda - \ln\Gamma(r) + (r-1)\ln x - \lambda x$

$\frac{\vartheta}{\vartheta\lambda}\ln f(x;\lambda) = \frac{r}{\lambda} - x$; $\frac{\vartheta}{\vartheta\lambda}\ln L(\lambda) = \frac{nr}{\lambda} - \sum_{i=1}^{n} x_i$. $\widehat{\lambda}^{MV}$ (estimation) solution annulant cette expression, donc $\widehat{\lambda}^{MV} = \frac{r}{\overline{X}}$ (estimateur).

Pour n grand $\widehat{\lambda}^{MV} \underset{approx}{\rightsquigarrow} \mathcal{N}(\lambda\,;\frac{1}{nI(\lambda)})$ selon la proposition 6.11. Calculons $I(\lambda)$:

$\frac{\vartheta^2}{\vartheta\lambda^2}\ln f(x;\lambda) = -\frac{r}{\lambda^2}$; $I(\lambda) = E\left[-\frac{\vartheta^2}{\vartheta\lambda^2}\ln f(X;\lambda)\right] = \frac{r}{\lambda^2}$,

d'où $\widehat{\lambda}^{MV} \underset{approx}{\rightsquigarrow} \mathcal{N}(\lambda\,;\frac{\lambda^2}{nr})$. Ce résultat sera utile pour établir un intervalle de confiance pour λ (voir chapitre 7).

Exercice 6.23

La densité de probabilité *a posteriori* de p sachant $(X_1,\cdots,X_n) = (x_1,\cdots,x_n)$ – ou en bref $\mathbf{X} = \mathbf{x}$ – est, en posant $s = \sum_{i=1}^{n} x_i$:

$$\pi_{p|\mathbf{X}=\mathbf{x}}(p) = \frac{p^s(1-p)^{n-s}\sqrt{p(1-p)}}{\int_0^1 p^s(1-p)^{n-s}\sqrt{p(1-p)}dp} = \frac{p^{s+\frac{1}{2}}(1-p)^{n-s+\frac{1}{2}}}{\int_0^1 p^{s+\frac{1}{2}}(1-p)^{n-s+\frac{1}{2}}dp}$$

qui est la densité de la loi $Beta(s+\frac{1}{2}, n-s+\frac{1}{2})$. Pour l'estimation bayésienne de p on prend sa moyenne $\frac{\Gamma(s+\frac{5}{2})}{\Gamma(n+4)}.\frac{\Gamma(n+3)}{\Gamma(s+\frac{3}{2})} = \frac{s+\frac{3}{2}}{n+3}$.

Pour une loi *a priori* $Beta(\alpha,\beta)$, on a une densité proportionnelle à $p^\alpha(1-p)^\beta$, avec $\alpha > -1, \beta > -1$, d'où une loi *a posteriori* $Beta(s+\alpha, n-s+\beta)$, de moyenne $\frac{s+\alpha+1}{n+2+\alpha+\beta}$.

Chapitre 7 : Estimation paramétrique par intervalle de confiance

Exercice 7.1

Pour μ la largeur est $2 \times 1,96 \frac{s}{\sqrt{n}}$.
Pour σ on a l'IC (voir note 7.5) :

$$IC_{0,95}(\sigma) = \left[\frac{\sqrt{(n-1)}s}{\sqrt{\chi^{2\ (n-1)}_{0,975}}} , \frac{\sqrt{(n-1)}s}{\sqrt{\chi^{2\ (n-1)}_{0,025}}} \right].$$

Pour n grand la loi $\chi^2(n-1)$, de moyenne $n-1$ et variance $2(n-1)$, est approchée par la loi $\mathcal{N}(n-1\,;2(n-1))$ (voir la remarque de la section 5.8.3). Donc :

$\chi^{2\ (n-1)}_{0,975} \simeq (n-1) + 1,96\sqrt{2(n-1)} = (n-1)\left[1 + 1,96\sqrt{\frac{2}{(n-1)}}\right]$

$\left[\chi^{2\ (n-1)}_{0,975}\right]^{-\frac{1}{2}} \simeq (n-1)^{-\frac{1}{2}}\left[1 + 1,96\sqrt{\frac{2}{(n-1)}}\right]^{-\frac{1}{2}} \simeq (n-1)^{-\frac{1}{2}}\left[1 - 1,96\sqrt{\frac{1}{2n}}\right]$

et de même par simples changements de signes :

$\left[\chi^{2\ (n-1)}_{0,025}\right]^{-\frac{1}{2}} \simeq (n-1)^{-\frac{1}{2}}\left[1 + 1,96\sqrt{\frac{1}{2n}}\right]$.

D'où $IC_{0,95}(\sigma) \simeq \left[s(1 - 1,96\sqrt{\frac{1}{2n}})\,,\ s(1 + 1,96\sqrt{\frac{1}{2n}})\right]$ dont la largeur est $2 \times 1,96\frac{s}{\sqrt{2n}}$ soit $\sqrt{2}$ fois plus petite que celle de l'IC sur μ.

Exercice 7.2

Pour un échantillon $(X_1, X_2, \cdots, X_n)$ de la loi $\mathcal{N}(0\,;\sigma^2)$, on a $\frac{\sum_{i=1}^n X_i^2}{\sigma^2} \rightsquigarrow \chi^2(n)$.
Donc :
$P(\chi^{2\ (n)}_{\frac{\alpha}{2}} < \frac{\sum_{i=1}^n X_i^2}{\sigma^2} < \chi^{2\ (n)}_{1-\frac{\alpha}{2}}) = 1-\alpha,\ P(\frac{\sum_{i=1}^n X_i^2}{\chi^{2\ (n)}_{1-\frac{\alpha}{2}}} < \sigma^2 < \frac{\sum_{i=1}^n X_i^2}{\chi^{2\ (n)}_{\frac{\alpha}{2}}}) = 1-\alpha,$
d'où $IC_{1-\alpha}(\sigma^2) = \left[\frac{\sum_{i=1}^n x_i^2}{\chi^{2\ (n)}_{1-\frac{\alpha}{2}}},\ \frac{\sum_{i=1}^n x_i^2}{\chi^{2\ (n)}_{\frac{\alpha}{2}}}\right]$.

Exercice 7.3

Pour $X \rightsquigarrow \mathcal{U}[0,\theta]$ on a la fonction de répartition $F_X(x;\theta) = \frac{x}{\theta}$ si $x \in [0,\theta]$. Donc pour un échantillon de taille n, $F_{X_{(n)}}(x;\theta) = [F_X(x;\theta)]^n = \left(\frac{x}{\theta}\right)^n$ si $x \in [0,\theta]$. Le quantile x_α de $X_{(n)}$ est tel que $\left(\frac{x_\alpha}{\theta}\right)^n = \alpha$ ou $x_\alpha = \theta\alpha^{\frac{1}{n}}$. Ainsi :

$$P\left(\theta(0,025)^{\frac{1}{n}} < X_{(n)} < \theta(0,975)^{\frac{1}{n}}\right) = 0,95$$

$$P\left(X_{(n)}(0,975)^{-\frac{1}{n}} < \theta < X_{(n)}(0,025)^{-\frac{1}{n}}\right) = 0,95$$

d'où $IC_{0,95}(\theta) = \left[x_{(n)}(0,975)^{-\frac{1}{n}},\ x_{(n)}(0,025)^{-\frac{1}{n}}\right]$.

Exercice 7.4

La densité de la loi $\chi^2(2n)$ est, au point x, $\frac{1}{2^n(n-1)!}x^{n-1}e^{-\frac{x}{2}}$ $(x > 0)$ et sa fonction de répartition est :

$$F_{\chi^2}(x;2n) = \int_0^x \frac{1}{2^n(n-1)!}t^{n-1}e^{-\frac{t}{2}}dt = \int_0^{\frac{x}{2}} \frac{1}{(n-1)!}t^{n-1}e^{-t}dt = J_{n-1}.$$

Intégrons par partie J_k pour établir une relation de récurrence :

$$J_k = -\int_0^{\frac{x}{2}} \frac{t^k}{k!}d(e^{-t}) = -\left[\frac{t^k}{k!}e^{-t}\right]_0^{\frac{x}{2}} + \int_0^{\frac{x}{2}} \frac{t^{k-1}e^{-t}}{(k-1)!}dt = -\frac{\left(\frac{x}{2}\right)^k e^{-\frac{x}{2}}}{k!} + J_{k-1},$$

avec $J_0 = 1 - e^{-\frac{x}{2}}$.

D'où $J_{n-1} = 1 - e^{-\frac{x}{2}} - \frac{\left(\frac{x}{2}\right)^2 e^{-\frac{x}{2}}}{2!} - \cdots \frac{\left(\frac{x}{2}\right)^{n-1} e^{-\frac{x}{2}}}{(n-1)!} = 1 - \sum_{k=0}^{n-1} \frac{e^{-\frac{x}{2}}\left(\frac{x}{2}\right)^k}{k!}$.

Remplaçons x par 2λ et $2x+2$, x entier, par $2n$ pour obtenir $F_{\chi^2}(2\lambda;2x+2) = 1 - \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!} = 1 - F_P(x;\lambda)$. Trouver λ tel que, pour x donné, $F_P(x;\lambda) = \alpha$ équivaut à prendre le quantile d'ordre $1-\alpha$ de la loi $\chi^2(2x+2)$ puis à le diviser par 2.

Exemple 7.6 : $T \rightsquigarrow P(7\lambda)$. Pour $x = 18$ et $\alpha = 0,025$ on prend pour 7λ le quantile $0,975$ de la loi $\chi^2(38)$, soit 56,9. Cela donne la borne supérieure de l'IC sur 7λ : $7\lambda_2 = \frac{56,9}{2} = 28,4$. Pour la borne inférieure, on obtient $x = 17$, $\alpha = 0,975$, $\chi^2_{0,025}(36) = 21,34$ et $7\lambda_1 \simeq 10,7$. Finalement, on a trouve bien le même intervalle pour λ : $[1,53\,;\,4,06]$.

Exercice 7.5

On a $T = \sum_{i=1}^{20} X_i \rightsquigarrow B(20\,;\,p)$. Il faut résoudre en p les deux équations (voir section 7.5, cas d'une loi discrète) :

$\sum_{x=0}^{7} \binom{20}{x} p^x (1-p)^{20-x} = 0,975$

$\sum_{x=0}^{8} \binom{20}{x} p^x (1-p)^{20-x} = 0,025$

Un logiciel de résolution d'équations donne pour la première $p_1 = 0,19$ et la deuxième $p_2 = 0,64$, qui sont les bornes de l'intervalle de confiance à 95%. L'intervalle asymptotique classique $\widehat{p} \pm 1,96 \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ de la section 7.4.5 donnerait, avec ici $\widehat{p} = 8/20$: $[0,185; 0,615]$. Cet intervalle est toutefois assez approximatif du fait que la règle de validité donnée $n\widehat{p}(1-\widehat{p}) > 12$ n'est pas vérifiée (ici $n\widehat{p}(1-\widehat{p}) = 4,8$).

Exercice 7.6

On sait que $\overline{X}$ est statistique exhaustive minimale. On a remarqué en section 6.8 que la loi *a posteriori* du paramètre inconnu sachant les valeurs prises par l'échantillon ne dépend que de cette statistique. On conditionnera donc simplement sur l'événement $(\overline{X} = \overline{x})$. Comme $\overline{X} \rightsquigarrow \mathcal{N}(\mu\,,\frac{\sigma^2}{n})$, sa densité est $f_{\overline{X}}(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2/n}\}$. La densité de la loi *a priori* de μ étant $\pi(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\{-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\}$, on a la loi *a posteriori* :

$$\pi_{\mu|\overline{X}=\overline{x}}(\mu) = c \, \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu)^2}{\sigma^2/n} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]\right\}$$

où c est la constante qui normalise à une densité. L'expression entre crochets s'écrit :

$$\mu^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) + \frac{\overline{x}^2}{\sigma^2/n} + \frac{\mu_0^2}{\sigma_0^2}$$

$$= \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\left[\mu - \frac{\frac{n\overline{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right]^2 + C^{te} = \frac{1}{\frac{\sigma_0^2\sigma^2/n}{\sigma_0^2+\sigma^2/n}}\left[\mu - \frac{\sigma_0^2\,\bar{x} + \sigma^2\mu_0/n}{\sigma_0^2 + \sigma^2/n}\right]^2 + C^{te},$$

expression qui met en évidence la moyenne et la variance de la loi *a posteriori* de μ, comme indiqué dans l'énoncé. On en déduit l'intervalle de probabilité $0,95$ pour μ (ou, en d'autres termes, son IC bayésien à 95%) :

$$\frac{\sigma_0^2\,\bar{x} + \sigma^2\mu_0/n}{\sigma_0^2 + \sigma^2/n} \pm 1,96\sqrt{\frac{\sigma_0^2\sigma^2/n}{\sigma_0^2 + \sigma^2/n}}.$$

Quand $n \to \infty$, on obtient $\overline{x}$ pour le premier terme et, pour l'expression sous le radical, $\frac{\sigma^2}{n+\sigma^2/\sigma_0^2} \sim \frac{\sigma^2}{n}$, soit l'IC classique.
Dans l'approche bayésienne, le centre de l'intervalle est une pondération entre $\overline{x}$ et μ_0 avec poids respectifs σ_0^2 et σ^2/n. La demi-largeur est plus petite car égale à $1,96\sqrt{\frac{\sigma^2}{n}}\sqrt{\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}}$.

Exercice 7.7

Dans les notations de la section 7.4.5, il suffit de montrer que $\sqrt{\frac{\widehat{P}_n(1-\widehat{P}_n)}{n}}$ tend vers 0 en probabilité, où $\widehat{P}_n = \frac{S_n}{n}$. Or selon la loi des grands nombres, $\widehat{P}_n$ converge presque sûrement vers p (voir section 5.8.2) et *a fortiori* $\widehat{P}_n$ converge en probabilité vers p. Ainsi $\sqrt{\widehat{P}_n(1-\widehat{P}_n)}$, en tant que fonction continue de $\widehat{P}_n$, converge en probabilité vers $\sqrt{p(1-p)}$, ce qui prouve le résultat.

Exercices appliqués

Exercice 7.8

On calcule $\overline{x} = 8,10$, $s^2 = 0,1018$, $s = 0,319$ et on lit $t_{0,975}^{(11)} = 2,201$. D'où :

$$IC_{0,95}(\mu) = 8,10 \pm 2,201\frac{0,319}{\sqrt{12}} = 8,10 \pm 0,20.$$

Exercice 7.9

On a $\overline{x} = 5,4$, $s = 3,1$ et on lit $t_{0,95}^{(499)} \simeq z_{0,95} = 1,645$. D'où :

$$IC_{0,90}(\mu) = 5,4 \pm 1,645\frac{3,1}{\sqrt{500}} = 5,4 \pm 0,23 \text{ jours.}$$

Pour le coût moyen, on obtient un IC à 90% en multipliant les bornes du précédent par 200, soit 1080 ± 46 euros.

Exercice 7.10

Pour les 200 sinistres sélectionnés, on a une valeur moyenne de 9 944 euros avec un écart-type de 1901. Avec $t_{0,975}^{(199)} \simeq 1,97$ on obtient, pour la valeur moyenne des sinistres en cours, l'IC à 95% : $9\,944 \pm 1,97\frac{1\,901}{\sqrt{200}}$ ou $9\,944 \pm 264,8$ euros.
Pour la valeur totale des sinistres en cours on a un IC à 95% de :

$11\,210 \times 9\,944 \pm 11\,210 \times 264,8 \simeq 111,5 \pm 3,0$ millions d'euros.

Exercice 7.11

Soit μ le nombre moyen de mots par page du livre, on a, avec $t_{0,975}^{(19)} = 2,093$:

$$IC_{0,95}(\mu) = 614 \pm 2,093\frac{26}{\sqrt{20}} = 614 \pm 12,2.$$

Pour le nombre total de mots, on obtient l'IC à 95% en multipliant par le nombre de pages, soit $97\,012 \pm 1\,923$ ou environ $[95\,100,\ 98\,900]$.

Exercice 7.12

Soit μ_1 et μ_2 les consommations moyennes avec carburant traditionnel et carburant nouveau respectivement. On cherche un IC sur $\mu_1 - \mu_2$. Sur les échantillons on a $n_1 = 10$, $\overline{x}_1 = 10,8$, $s_1 = 0,21$, $n_2 = 10$, $\overline{x}_2 = 10,3$, $s_2 = 0,18$. On peut appliquer la formule sous l'hypothèse de variances égales au vu des tailles d'échantillons et des écarts-types observés (voir section 7.4.3). On calcule d'abord la variance empirique pondérée : $s_p^2 = \frac{1}{18}(9 \times (0,21)^2 + 9 \times (0,18)^2) = 0,03825$, d'où $s_p = 0,196$.
Avec $t_{0,95}^{(18)} = 1,734$ on obtient :

$$IC_{0,95}(\mu_1 - \mu_2) = (10,8 - 10,3) \pm 1,734 \times 0,196\sqrt{\frac{1}{10} + \frac{1}{10}} \simeq 0,5 \pm 0,15.$$

Le gain peut être estimé entre 0,35 et 0,65 litres.

Exercice 7.13

On a $\widehat{p} = 45/400 = 0,1125$. Comme $n\widehat{p}(1-\widehat{p}) = 39,9 > 12$ nous appliquons l'approximation gaussienne pour la proportion de pièces défectueuses. Avec $z_{0,995} = 2,57$ on a :

$$IC_{0,99}(p) = 0,1125 \pm 2,57\sqrt{\frac{0,1125 \times 0,8875}{400}} = 0,1125 \pm 0,0406,$$

ce qui donne pour le nombre total de pièces défectueuses dans le stock : $1\,125 \pm 406$ soit entre 720 et 1 530 pièces environ.

Exercice 7.14

On a $\widehat{p} = 0,20$. Comme $n\widehat{p}(1-\widehat{p}) = 240 > 12$, nous appliquons l'approximation gaussienne pour la proportion de personnes prévoyant d'acheter une voiture dans les douze prochains mois. Avec $z_{0,975} = 1,96$ on a :

$$IC_{0,95}(p) = 0,20 \pm 1,96\sqrt{\frac{0,20 \times 0,80}{1500}} = 0,20 \pm 0,032,$$

soit entre 17% et 23% environ.

Exercice 7.15

Soit p_1 et p_2 les proportions de pièces défectueuses produites par le premier et le deuxième procédé respectivement. On cherche un IC sur $p_1 - p_2$. Sur les échantillons on a $n_1 = 1000$, $\widehat{p}_1 = 86/1000 = 0,086$, $n_2 = 800$, $\widehat{p}_2 = 92/800 = 0,115$. Comme $n\widehat{p}_1(1-\widehat{p}_1) = 78,6 > 12$ et $n\widehat{p}_2(1-\widehat{p}_2) = 81,4 > 12$ nous appliquons l'approximation gaussienne de la section 7.4.6. Avec $z_{0,975} = 1,96$ on a :

$$\begin{aligned} IC_{0,95}(p_1 - p_2) &= (0,086 - 0,115) \pm 1,96\sqrt{\frac{0,086 \times 0,914}{1000} + \frac{0,115 \times 0,885}{800}} \\ &= -0,029 \pm 0,028, \end{aligned}$$

soit entre $-0,057$ et $-0,001$. Cet intervalle semble indiquer que p_1 est inférieur à p_2.

Exercice 7.16

Pour les 50 jours, on a observé un nombre total d'accidents égal à $\sum_{i=1}^{50} x_i = 0 \times 21 + 1 \times 18 + 2 \times 7 + 3 \times 3 + 4 \times 1 = 45$ d'où une moyenne d'accident par jour observée de $\overline{x} = 0,90$. On applique l'approximation gaussienne développée dans l'exemple 7.3. Avec $z_{0,975} = 1,96$ on a :

$$IC_{0,95}(\lambda) = 0,90 \pm 1,96\sqrt{\frac{0,90}{50}} = 0,90 \pm 0,26,$$

soit un nombre moyen d'accident par jour entre 0,64 et 1,16.
Notons que cet IC tient compte des spécificités du modèle de Poisson, à savoir que moyenne et variance sont égales à λ.

Exercice 7.17

On calcule $s^2 = 0,00461$, $s = 0,0679$. Calculons d'abord l'IC pour la variance σ^2 du taux (voir section 7.4.2), avec $\chi^{2\ (9)}_{0,025} = 2,700$ et $\chi^{2\ (9)}_{0,975} = 19,023$:

$$IC_{0,95}(\sigma^2) = [\frac{9 \times 0,00461}{19,023},\ \frac{9 \times 0,00461}{2,700}] = [0,00281\,,\ 0,01537]$$

soit $IC_{0,95}(\sigma) = [0,047\,,\ 0,124]$ ou, environ, $[0,05\,,\ 0,12]$.

Chapitre 8 : Estimation non paramétrique et estimation fonctionnelle

Exercice 8.1

Dans EXCEL on génère en première colonne 200 nombres au hasard dans [0,1] avec la fonction ALEA(). En deuxième colonne on applique à la première colonne la fonction EXP(LOI.NORMALE.STANDARD.INVERSE(-)) ou directement LOI.LOGNORMALE.INVERSE(- ; 0 ; 1), pour obtenir 200 observations issues de la loi $\mathcal{LN}(0\,;1)$ selon le principe défini en section 4.3.

Nous avons effectué cette opération et trouvé une médiane des 200 observations égale à 1,073 qui est une estimation ponctuelle de la vraie médiane $\widetilde{\mu}$ laquelle est ici égale à $e^0 = 1$. En effet, si ζ_q est le quantile d'ordre q de la loi $\mathcal{N}(0\,;1)$, e^{ζ_q} est le quantile d'ordre q de la loi $\mathcal{LN}(0\,;1)$ puisque $P(Z \leq q) = P(e^Z \leq e^q) = q$ où $Z \rightsquigarrow \mathcal{N}(0\,;1)$ et $e^Z \rightsquigarrow \mathcal{LN}(0\,;1)$.

Comme vu en section 8.2 on obtient un IC à 95% pour $\widetilde{\mu}$ avec les l_1-ième et $l_2 + 1$-ième statistiques d'ordre réalisées dans l'échantillon, où l_1 est tel que $P(\widetilde{N} \geq l_1) \geq 0,975$ et $l_2 = n - l_1$, avec $\widetilde{N} \rightsquigarrow B(200\,;\,0,5)$. En calculant les probabilités de cette loi dans EXCEL pour l'ensemble des valeurs possibles on trouve que $P(86 \leq \widetilde{N}) = 0,980$ et $P(87 \leq \widetilde{N}) = 0,972$ donc $l_1 = 86$ et $l_2 = 114$. Dans notre échantillon, les 86-ième et 115-ième statistiques d'ordre ont pris, respectivement, les valeurs 0,923 et 1,230 d'où $IC_{0,95}(\widetilde{\mu}) = [0{,}923;1{,}230]$. Cet intervalle couvre bien la vraie valeur.

Pour $x_{0,90}$, le quantile d'ordre 0,90, on recourt de même à la v.a. $\widetilde{N}^*$, nombre d'observations inférieures ou égales à $x_{0,90}$ qui suit une loi $B(200\,;\,0,90)$. On doit chercher l_1^* et l_2^* tels que $P(l_1^* \leq \widetilde{N}^* \leq l_2^*) \geq 0,95$. On trouve, en calculant les probabilités de cette loi dans EXCEL, $P(171 \leq \widetilde{N}^*) = 0,984$ et $P(172 \leq \widetilde{N}^*) = 0,973$ donc on prendra $l_1^* = 171$. Comme $P(\widetilde{N}^* \leq 187) = 0,968$ on a $P(171 \leq \widetilde{N}^* \leq 187) = 0,968 - 0,016 = 0,952$. Dans notre échantillon les 171-ième et 188-ième statistiques d'ordre ont pris, respectivement, les valeurs 3,29 et 4,48 d'où $IC_{0,95}(\widetilde{\mu}) = [3{,}29;4{,}48]$. Cet intervalle couvre bien la vraie valeur égale à $e^{1,28} = 3{,}60$. L'estimation ponctuelle était donnée par le quantile empirique d'ordre 0,90 soit 3,94 pour notre échantillon.

Exercice 8.2

Pour les 12 valeurs données, on obtient $\overline{x} = 1,10$ et $s = 1,32$.
On calcule les valeurs $s_{-i} = \sqrt{\frac{1}{10}\sum_{j=1,j\neq i}^{12}(x_i - \overline{x}_{-i})^2}$ et on obtient :

1,34 1,34 1,34 1,34 1,36 1,37 1,38 1,38 1,38 1,36 1,36 0,71.

Puis on calcule les pseudo-valeurs $s_{*i} = 12s - 11s_{-i}$ et on obtient :

1,05 1,05 1,05 1,05 0,82 0,77 0,62 0,61 0,63 0,89 0,89 8,05.

La moyenne des pseudo-valeurs est égale à 1,46 ce qui donne l'estimation par jackknife de σ. On note qu'elle est assez différente de l'estimation usuelle par s =1,32 ce qui s'explique par l'incidence forte de la dernière valeur pour un nombre faible d'observations.

Comme l'écart-type des pseudo-valeurs est s_{JK} =2,08 on obtient :

$IC_{0,95}(\sigma) = [1{,}46 - t^{11}_{0,975}\frac{2{,}08}{\sqrt{12}}, 1,46 + t^{11}_{0,975}\frac{2{,}08}{\sqrt{12}}]$

soit, avec $t^{11}_{0,975}$ =2,201, l'intervalle [0,14 ; 2,78] qui est assez large en raison du faible nombre d'observations.
Note : Les résultats ont été calculés avec la précision d'EXCEL mais sont indiqués arrondis à la deuxième décimale.

Exercice 8.3

On a $\widetilde{S}^2_n = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^2$ et, en omettant l'observation i :

$$\widetilde{S}^2_{-i} = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}(X_j - \overline{X}_{-i})^2 = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}(X_j - \overline{X})^2 - (\overline{X}_{-i} - \overline{X})^2$$

soit pour la pseudo-valeur correspondante :

$$\begin{aligned}\widetilde{S}^2_{*i} &= n\widetilde{S}^2_n - (n-1)\widetilde{S}^2_{-i}\\ &= \sum_{j=1}^{n}(X_j - \overline{X})^2 - \sum_{j=1,j\neq i}^{n}(X_j - \overline{X})^2 + (n-1)(\overline{X}_{-i} - \overline{X})^2\\ &= (X_i - \overline{X})^2 + (n-1)(\overline{X}_{-i} - \overline{X})^2.\end{aligned}$$

Or :

$$\begin{aligned}\overline{X}_{-i} - \overline{X} &= \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}X_j - \frac{1}{n}\sum_{j=1}^{n}X_j = \left(\frac{1}{n-1} - \frac{1}{n}\right)\sum_{j=1}^{n}X_j - \frac{1}{n-1}X_i\\ &= \frac{1}{n(n-1)}\sum_{j=1}^{n}X_j - \frac{1}{n-1}X_i = \frac{1}{n-1}(\overline{X} - X_i),\end{aligned}$$

d'où :

$$\widetilde{S}^2_{*i} = (X_i - \overline{X})^2 + \frac{1}{n-1}(\overline{X} - X_i)^2 = \frac{n}{n-1}(X_i - \overline{X})^2.$$

Ainsi l'estimateur du Jackknife qui est la moyenne des $\widetilde{S}^2_{*i}$ donne S^2_n qui est sans biais. Cela corrobore la proposition 8.2 : le biais de $\widetilde{S}^2_n$ qui est $-\frac{\sigma^2}{n}$ est éliminé par la procédure Jackknife.

Exercice 8.4

On a $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ et $\overline{X}_{-i} = \frac{1}{n-1}\sum_{j=1, j\neq i}^n X_j$. Donc :

$$\overline{X}_{*i} = n\overline{X} - (n-1)\overline{X}_{-i} = \sum_{j=1}^n X_j - \sum_{j=1, j\neq i}^n X_j = X_i.$$

Les pseudo-valeurs sont ainsi identiques aux valeurs mêmes et l'estimateur du jackknife reste $\overline{X}$.

Exercice 8.5

Considérons la double suite $\cdots, a_0 - 2h, a_0 - h, a_0, a_0 + h, a_0 + 2h, \cdots$ qui définit une grille d'intervalles de largeur h pour l'histogramme, pour laquelle nous considérerons le point a_0 comme point de positionnement. Soit $\widehat{f}_n(x; a_0)$ la valeur de l'histogramme au point x. Nous pouvons écrire $\widehat{f}_n(x; a_0)$ en exprimant le comptage des x_i situés dans le même intervalle que x (pour lever l'ambiguïté aux limites des intervalles, nous prendrons des intervalles ouverts à droite), ce qui donne la double somme comme suit :

$$\begin{aligned}\widehat{f}_n(x; a_0) &= \frac{1}{nh}\sum_{i=1}^n \sum_{k\in\mathbb{Z}} I_{[a_0+kh, a_0+(k+1)h[}(x_i)\, I_{[a_0+kh, a_0+(k+1)h[}(x) \\ &= \frac{1}{nh}\sum_{i=1}^n \sum_{k\in\mathbb{Z}} I_{[0,1[}(\frac{x_i - a_0}{h} - k)\, I_{[0,1[}(\frac{x - a_0}{h} - k).\end{aligned}$$

Nous considérons maintenant $\overline{\widehat{f}}_n(x; a_0)$ correspondant à la valeur moyenne obtenue en faisant glisser uniformément la grille du positionnement a_0 à $a_0 + h$:

$$\overline{\widehat{f}}_n(x; a_0) = \frac{1}{nh^2}\int_{a_0}^{a_0+h} \sum_{i=1}^n \sum_{k\in\mathbb{Z}} I_{[0,1[}(\frac{x_i - t}{h} - k)\, I_{[0,1[}(\frac{x - t)}{h} - k)dt,$$

soit, en posant $u = \frac{t-a_0}{h}$, $z_i = \frac{x_i-a_0}{h}$, $z = \frac{x-a_0}{h}$ et en permutant l'intégration et la première sommation :

$$\overline{\widehat{f}}_n(x;a_0) = \frac{1}{nh}\sum_{i=1}^{n}\int_0^1 \sum_{k\in\mathbb{Z}} I_{[0,1[}(z_i-k-u)I_{[0,1[}(z-k-u)du.$$

En remarquant que $I_{[0,1[}(z_i-k-u)$ est égal à 1, si et seulement si $k=[z_i]$ (la partie entière de z_i) quand $u \in [0, z_i-[z_i]]$ et si et seulement $k=[z_i]-1$ quand $u \in [z_i-[z_i],1]$, on a :

$$\overline{\widehat{f}}_n(x;a_0) = \frac{1}{nh}\sum_{i=1}^{n}\left[\int_0^{z_i-[z_i]} I_{[0,1[}(z-[z_i]-u)du + \int_{z_i-[z_i]}^1 I_{[0,1[}(z-[z_i]+1-u)du\right].$$

En effectuant le changement de variable $v = z-[z_i]-u$ dans la première intégrale et $v = z-[z_i]+1-u$, on obtient :

$$\overline{\widehat{f}}_n(x;a_0) = \frac{1}{nh}\sum_{i=1}^{n}\int_{z-z_i}^{z-z_i+1} I_{[0,1[}(v)dv.$$

En examinant les différentes positions de $[0,1]$ par rapport à $[z-z_i, z-z_i+1]$, on arrive finalement à :

$$\begin{aligned}\overline{\widehat{f}}_n(x;a_0) &= \frac{1}{nh}\sum_{i=1}^{n}\left[1-|z-z_i|\right] I_{[0,1[}(|z-z_i|)\\ &= \frac{1}{nh}\sum_{i=1}^{n}K(z-z_i) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)\end{aligned}$$

où $K(u) = (1-|u|)I_{[0,1[}(|u|)$ est le noyau triangulaire.
Comme on s'y attendait, $\overline{\widehat{f}}_n(x;a_0)$ ne dépend pas de l'origine a_0.

Exercice 8.6

Dans EXCEL, on génère en première colonne 50 nombres au hasard dans [0,1] avec la fonction ALEA(). En deuxième colonne on applique à la première colonne la fonction LOI.NORMALE.STANDARD.INVERSE(-). Puis on applique la fonction biweight avec $h=1$ soit $\frac{15}{16}(1-x^2)^2$ pour les valeurs $x \in [-1,1]$ et 0 sinon. Alors la moyenne de ces poids est l'estimation pour $f(0) = \frac{1}{\sqrt{2\pi}} \simeq$0,399. L'espérance mathématique de cet estimateur est (voir section 8.5.2) :

$$\frac{1}{h}\int_{\mathbb{R}} K(\frac{x-t}{h})f(t)dt = \int_{-1}^{+1}\frac{15}{16}(1-x^2)^2\,\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \simeq 0{,}369\,,$$

valeur obtenue via un logiciel de calcul mathématique. Le biais est 0,369−0,399 = 0,030. Il est évidemment négatif puisqu'on pondère des observations prises au voisinage d'un maximum.
Sur une simulation de 50 observations on doit trouver une estimation avec un écart à la vraie valeur pas trop éloigné de 0,03 (on peut éventuellement augmenter la taille de l'échantillon pour le vérifier plus précisément).

Exercice 8.7

On procède comme à l'exercice précédent à ceci près que la fonction de poids est $\frac{15}{16}(1-(\frac{x}{h})^2)^2$ pour les $x \in [-h,h]$, avec les différentes valeurs de h proposées.

La valeur de h optimale asymptotiquement **en un point** x est donnée en section 8.5.2 :

$$h_{opt} = \left[\frac{f(x) \int_{\mathbb{R}} [K(u)]^2 \, du}{[f''(x)]^2 \left[\int_{\mathbb{R}} u^2 K(u) du\right]^2} \right]^{1/5} n^{-1/5}.$$

Ici on a $x=0$, $f(0) = \frac{1}{\sqrt{2\pi}} \simeq 0{,}399$, $f''(0) = -1$ et :

$$\int_{\mathbb{R}} [K(u)]^2 \, du = \left(\frac{15}{16}\right)^2 \int_{-1}^{+1} (1-u^2)^4 du = \left(\frac{15}{16}\right)^2 \int_{-1}^{+1} (1 - 4u^2 + 6u^4 - 4u^6 + u^8) du$$
$$= 2\left(\frac{15}{16}\right)^2 \left[1 - \frac{4}{3} + \frac{6}{5} - \frac{4}{7} + \frac{1}{9}\right] \simeq 0{,}7143 \ ,$$

$$\int_{\mathbb{R}} u^2 K(u) du = \frac{15}{16} \int_{-1}^{+1} (u^2 - 2u^4 + u^6) du = 2\frac{15}{16}\left[\frac{1}{3} - \frac{2}{5} + \frac{1}{7}\right] \simeq 0{,}1429.$$

d'où finalement h_{opt} =1,694 $n^{-\frac{1}{5}}$.
Pour $n = 50$ on trouve h_{opt} =0,775. Observez-vous l'estimation la plus proche de $\frac{1}{\sqrt{2\pi}}$ ≃0,399 pour h =0,75? Augmentez éventuellement la taille de l'échantillon pour vérifier empiriquement que la meilleure estimation est celle obtenue avec la valeur de h la plus proche de h_{opt}. Par exemple, pour $n = 500$, h_{opt} = 0,49. Notons que la largeur de fenêtre reste assez large même pour une taille d'échantillon assez élevée.

Exercice 8.8

Rappelons l'expression asymptotique de l'eqim donnée en section 8.5.2.

$$eqim\left(\widehat{f}_n\right) \sim \frac{h^4}{4} \int_{\mathbb{R}} [f''(x)]^2 dx \ \left[\int_{\mathbb{R}} u^2 K(u) du\right]^2 + \frac{1}{nh} \int_{\mathbb{R}} [K(u)]^2 \, du$$

En la dérivant par rapport à h, on obtient une expression qui s'annule pour :

$$h_{opt} = \left[\frac{\int_{\mathbb{R}} [K(u)]^2 \, du}{\int_{\mathbb{R}} [f''(x)]^2 dx \left[\int_{\mathbb{R}} u^2 K(u) du \right]^2} \right]^{1/5} n^{-1/5}.$$

et en remplaçant dans l'eqim :

$$eqim \left(\widehat{f}_n \right)_{opt} = \frac{5}{4} \left[\int_{\mathbb{R}} [K(u)]^2 \, du \right]^{\frac{4}{5}} \left[\int_{\mathbb{R}} u^2 K(u) du \right]^{\frac{2}{5}} \left[\int_{\mathbb{R}} [f''(x)]^2 dx \right]^{\frac{1}{5}} n^{-\frac{4}{5}}.$$

Le facteur dépendant uniquement du noyau que l'on peut souhaiter minimiser est $\nu(K) = \left[\int_{\mathbb{R}} [K(u)]^2 \, du \right]^{\frac{4}{5}} \left[\int_{\mathbb{R}} u^2 K(u) du \right]^{\frac{2}{5}}$.
Pour le noyau biweight on a calculé à l'exercice précédent $\int_{\mathbb{R}} [K(u)]^2 \, du = 0{,}7143$ et $\int_{\mathbb{R}} u^2 K(u) du = 0{,}1429$ ce qui donne $\nu(K) = 0{,}351$. Pour le noyau de Rosenblatt on trouve de même $\nu(K) = 0{,}369$ et pour le meilleur noyau, celui d'Epanechnikov, $\nu(K) = 0{,}349$. On constate qu'en termes d'approximation asymptotique les différences sont faibles. En particulier le noyau biweight est très proche de celui d'Epanechnikov, avec l'avantage d'être partout dérivable.

Exercice 8.9

On peut simplement appliquer les résultats de l'exercice précédent concernant le noyau biweight :

$$h = \left[\frac{0{,}7143}{\left(\frac{3\sigma^5}{8\sqrt{\pi}} \right) (0{,}1429)^2} \right]^{\frac{1}{5}} n^{-\frac{1}{5}} \simeq 2{,}78 \, \sigma n^{-\frac{1}{5}}.$$

$$\begin{aligned} eqim \left(\widehat{f}_n \right)_{opt} &= \frac{5}{4} \nu(K) \left[\int_{\mathbb{R}} [f''(x)]^2 dx \right]^{\frac{1}{5}} n^{-\frac{4}{5}} \\ &= \frac{5}{4} (0{,}351) \left(\frac{3\sigma^5}{8\sqrt{\pi}} \right)^{\frac{1}{5}} n^{-\frac{4}{5}} \simeq 0{,}321 \, \sigma^{-1} n^{-\frac{4}{5}}. \end{aligned}$$

En intégrant $eqm \left(\widehat{f}_n(x) \right) \sim \frac{h^2}{12} [f'(x)]^2 + f(x)/(nh)$ donnée en section 8.5.2 pour l'histogramme au point x, on obtient :

$$eqim(\widehat{f}_n) = \int_{\mathbb{R}} eqm \left(\widehat{f}_n(x) \right) dx \sim \frac{h^2}{12} \int_{\mathbb{R}} [f'(x)]^2 \, dx + \frac{1}{nh}.$$

La dérivée par rapport à h de cette expression s'annule pour :

$$h_{opt} = 6^{1/3} \left[\int_{\mathbb{R}} [f'(x)]^2 \, dx \right]^{-1/3} n^{-1/3},$$

ce qui donne pour la loi $\mathcal{N}(\mu \, , \sigma^2)$:

$$h_{opt} = (24\sqrt{\pi})^{\frac{1}{3}}\,\sigma\, n^{-1/3} \simeq 3{,}49\,\sigma n^{-1/3}.$$

En remplaçant dans l'expression de $eqim(\widehat{f}_n)$ ci-dessus on vérifie aisément que $eqim(\widehat{f}_n)_{opt} \simeq 0{,}430\,\sigma^{-1} n^{-\frac{2}{3}}$.
Pour $n = 500$, par exemple, cette erreur vaut $6{,}83\times10^{-3}\sigma^{-1}$ contre $2{,}22\times10^{-3}\sigma^{-1}$ pour le biweight, soit environ trois fois plus.

Exercice 8.10

Pour $\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x-x_i}{h}\right)$, on a :

$$\begin{aligned} E\left[\widehat{F}_n(x)\right] &= \frac{1}{n}\sum_{i=1}^{n} E\left[H\left(\frac{x-X_i}{h}\right)\right] = E\left[H\left(\frac{x-X}{h}\right)\right] = \int_{-\infty}^{+\infty} H\left(\frac{x-t}{h}\right) f(t)dt, \\ &= \int_{-\infty}^{+\infty} H(u)\, h\, f(x-uh)du = \int_{-\infty}^{+\infty} H(u)\, d(-F(x-uh)) \\ &= [-H(u)F(x-uh]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} K(u)\, F(x-uh)du. \end{aligned}$$

car $H'(u) = K(u)$. Le premier terme est nul et, pour le deuxième, développons $F(x-uh)$ au voisinage de x :

$$\begin{aligned} E\left[\widehat{F}_n(x)\right] &= \int_{-\infty}^{+\infty} K(u)\left[F(x) - uhf(x) + \frac{u^2h^2}{2}f'(x) + o(h^2)\right]du \\ &= F(x) + \frac{h^2}{2}f'(x)\int_{-\infty}^{+\infty} u^2K(u)du + o(h^2) \end{aligned}$$

en utilisant la symétrie du noyau. Notons qu'en raison de cette propriété on peut remplacer $o(h^2)$ par $o(h^3)$.

Chapitre 9 : Tests d'hypothèses paramétriques

Exercice 9.1

La fonction de vraisemblance est $L(\lambda) = \prod_{i=1}^n f(x_i;\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$. Le rapport de vraisemblance (RV) est :

$$\frac{L(\frac{1}{2})}{L(1)} = \left(\frac{1}{2}\right)^n \exp\{(1-\frac{1}{2})\sum_{i=1}^n x_i\} = \left(\frac{1}{2}\right)^n \exp\{\frac{1}{2}\sum_{i=1}^n x_i\}.$$

On rejette $H_0 : \lambda = 1/2$ (vs. $H_1 : \lambda = 1$) si $\frac{L(\frac{1}{2})}{L(1)} < k_\alpha$, ce qui équivaut à $\sum_{i=1}^n x_i < k'_\alpha$. Sous H_0 $\sum_{i=1}^n X_i \rightsquigarrow \Gamma(n, \frac{1}{2}) \equiv \chi^2(2n)$.
Comme $P(\sum_{i=1}^n X_i < \chi^{2\,(2n)}_{0,05} | \lambda = \frac{1}{2}) = 0,05$ on rejette H_0 au niveau 0,05 si $\sum_{i=1}^n x_i < \chi^{2\,(2n)}_{0,05}$.

Exercice 9.2

a) Dans un processus de Bernoulli $X_1 \rightsquigarrow \mathcal{G}(p)$ est le nombre d'échecs avant le premier succès. Les v.a. X_1 et X_2 étant indépendantes, $X_1 + X_2$ peut être vue comme le nombre d'échecs avant le deuxième succès et ainsi de suite $X_1 + X_2 + \cdots + X_n$ peut-être vue comme le nombre d'échecs avant le n-ième succès et ainsi $\sum_{i=1}^n X_i \rightsquigarrow \mathcal{BN}(n,p)$.
b) La fonction de vraisemblance de p est $L(p) = p^n(1-p)^{\sum_{i=1}^n x_i}$, n étant connu. Pour le test considéré le RV est :

$$\frac{L\left(\frac{1}{3}\right)}{L\left(\frac{2}{3}\right)} = \left(\frac{1}{2}\right)^n 2^{\sum_{i=1}^n x_i}$$

et $\frac{L\left(\frac{1}{3}\right)}{L\left(\frac{2}{3}\right)} < k_\alpha$ équivaut à $\sum_{i=1}^n x_i < k'_\alpha$.
c) $T = \sum_{i=1}^n X_i \rightsquigarrow \mathcal{BN}(4, \frac{1}{3})$ sous H_0. Alors $P(T=0) = \binom{3}{0}\left(\frac{1}{3}\right)^4 (\frac{2}{3})^0 = 0,0123$, $P(T=1) = \binom{4}{1}\left(\frac{1}{3}\right)^4 (\frac{2}{3}) = 0,0329$ et $P(T=2) = \binom{5}{2}\left(\frac{1}{3}\right)^4 (\frac{2}{3})^2 = 0,0549$. On a donc un test conservateur de niveau 0,05 en rejetant H_0 si $\sum_{i=1}^n x_i \leq 1$.

d) La puissance du test est donnée par $P(T=1|\, p=\frac{2}{3}) = (\frac{2}{3})^4 + 4(\frac{2}{3})^4\left(\frac{1}{3}\right) = 0,461$.

Exercice 9.3

Le risque de première espèce est $\alpha = P(\overline{X} > 5 + \frac{1}{\sqrt{n}})$ avec $\overline{X} \rightsquigarrow \mathcal{N}(5\,;\frac{1}{n})$. Donc $\alpha = P(\frac{\overline{X}-5}{1/\sqrt{n}} > 1) = P(Z > 1)$ où $Z \rightsquigarrow \mathcal{N}(0\,;1)$, d'où $\alpha = 0,159$.
Pour μ quelconque $\overline{X} \rightsquigarrow \mathcal{N}(\mu\,,\frac{1}{n})$ et :

$$P(\overline{X} > 5 + \frac{1}{\sqrt{n}}) = P(\frac{\overline{X}-\mu}{1/\sqrt{n}} > \frac{5-\mu}{1/\sqrt{n}} + 1) = P(Z > \sqrt{n}(5-\mu) + 1)$$

et la fonction puissance, définie pour $\mu > 5$, est $h(\mu) = 1 - \Phi(\sqrt{n}(5-\mu)+1)$ où Φ est la fonction de répartition de la loi normale centrée-réduite. Cette fonction croît de 0,159 à 1.

Exercice 9.4

Écrivons la fonction de densité $f(x;\theta) = \theta a^\theta I_{[a,+\infty[}(x) \exp\{-(\theta+1)\ln x\}$, .soit la forme exponentielle avec $c(\theta) = -(\theta+1)$. Or la proposition 9.5 indique que le RV est monotone si la fonction $c(\theta)$ est monotone. Pour $\theta < \theta'$, $L(\theta)/L(\theta') = (\theta a^\theta/\theta' a^{\theta'})^n \exp\{(\theta'-\theta)\sum_{i=1}^n \ln x_i\}$ croît en fonction de la statistique exhaustive minimale (ou plutôt sa réalisation) $\sum_{i=1}^n \ln x_i$.

Pour $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$ le test UPP consiste à rejeter H_0 si $\sum_{i=1}^n \ln x_i > k_\alpha$ (cas N° 3 de la section 9.4.2).

Application :
L'existence de la moyenne $\frac{\theta}{\theta-1}$suppose que $\theta > 1$. L'hypothèse $\frac{\theta}{\theta-1} \leq 2$ équivaut à $\theta \geq 2$. Or le test UPP pour $H_0 : \theta \geq 2$ vs. $H_1 : 1 < \theta < 2$ consiste à rejeter H_0 si $\sum_{i=1}^n \ln x_i > k_\alpha$. La fonction de répartition de X suivant la loi de Pareto avec $a = 1$ étant, au point x, $1 - x^{-\theta}$ pour $x \geq 1$, celle de $\ln X$ au point x est $P(\ln X \leq x) = P(X \leq e^x) = 1 - e^{-\theta x}$ si $e^x \geq 1$, soit $x > 0$. Donc $\ln X \rightsquigarrow \mathcal{E}(\theta)$ et $\sum_{i=1}^n \ln X_i \rightsquigarrow \Gamma(n,\theta)$.
Sous H_0 $\sum_{i=1}^n \ln X_i \rightsquigarrow \Gamma(n,2)$. Soit $\gamma_{0,95}$ le quantile d'ordre 0,95 de cette loi, alors le test UPP consiste à rejeter H_0 au niveau $\alpha = 0,05$ si $\sum_{i=1}^n \ln x_i > \gamma_{0,95}$.

Exercice 9.5

On a $f(x;\theta) = \frac{1}{\theta} I_{[0,\theta]}(x)$ et :

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i) = \frac{1}{\theta^n} I_{[0,\theta]}(x_{(n)}) I_{[0,+\infty]}(x_{(1)}).$$

Pour $0 < \theta < \theta'$, $L(\theta)/L(\theta')$ est nul pour $x_{(n)} > \theta$ et vaut $(\theta'/\theta)^n > 0$ pour $0 \leq x_{(n)} \leq \theta$ (le RV n'est pas défini pour $x_{(n)} < 0$). Donc le RV est non croissant en fonction de $x_{(n)}$.

Soit, par exemple, le test $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$ (cas N° 2 de la section 9.4.2), alors le test UPP consiste à rejeter H_0 si $x_{(n)} > k$. Pour un niveau α on choisit k tel que, sous H_0, $P(X_{(n)} > k) = \alpha$. Or, sous H_0, la fonction de répartition de $X_{(n)}$ au point x est $\left(\frac{x}{\theta_0}\right)^n$, donc $P(X_{(n)} > k) = 1 - \left(\frac{k}{\theta_0}\right)^n$. En résolvant $1 - \left(\frac{k}{\theta_0}\right)^n = \alpha$, on trouve $k = \theta_0(1-\alpha)^{1/n}$.

Exercice 9.6

On est ici dans une situation où l'on fait une seule observation. X est donc statistique exhaustive minimale. Pour une réalisation x, on a :

$$L(M) = \binom{M}{x}\binom{N-M}{n-x} \Big/ \binom{N}{n}$$

si M entier dans $[n, N]$ et 0 sinon (nous envisageons le cas réaliste où M et $N - M$ sont supérieurs à n sinon il y a lieu de tenir compte de toutes les situations autres). D'où :

$$\frac{L(M+1)}{L(M)} = \frac{\binom{M+1}{x}\binom{N-M-1}{n-x}}{\binom{M}{x}\binom{N-M}{n-x}}$$
$$= \frac{(M+1)!(M-x)!(N-M-1)!(N-M-n+x)!}{(M+1-x)!M!(N-M-1-n+x)!(N-M)!} = \frac{(M+1)(N-M-n+x)}{(M+1-x)(N-M)}$$

qui est une fonction croissante de x. Cela est vrai également pour $\frac{L(M+2)}{L(M)} = \frac{L(M+2)}{L(M+1)}\frac{L(M+1)}{L(M)}$ comme produit de deux fonctions croissantes positives et, de proche en proche, pour $\frac{L(M')}{L(M)}$ avec $M' > M$. Notons que les résultats établis en section 9.4.2 sont fondés sur le rapport $\frac{L(\theta)}{L(\theta')}$ avec $\theta < \theta'$, lequel est donc ici décroissant.

Pour tester $H_0 : M \geq M_0$ vs. $H_1 : M < M_0$ nous sommes dans le cas 4 décrit en section 9.4.2 et le test UPP consiste à rejeter H_0 si $x < k$, ce qui est intuitif. Pour un risque de première espèce α choisi, soit il existe − cas très peu vraisemblable − un entier c_α tel que, pour $M = M_0$, $P(X \leq c_\alpha) = \alpha$ et on a un test exactement de niveau α en rejetant si $x \leq c_\alpha$, soit le quantile d'ordre α

est une interpolation entre un entier c et $c+1$, auquel cas on devra rejeter de façon conservatrice pour $x \leq c$, ce qui équivaut encore à $x \leq c_\alpha$ avec la valeur non entière d'interpolation.

Exercice 9.7

Soit le rejet de H_0 pour une observation n'appartenant pas à $[c_1, c_2]$. On a pour un λ donné :

$P_\lambda(X \notin [c_1, c_2]) = P_\lambda(X < c_1) + P_\lambda(X > c_2) = 1 - e^{-\lambda c_1} + e^{-\lambda c_2}$. Il s'agit de résoudre en (c_1, c_2) le système suivant :

$$\begin{cases} 1 - e^{-\lambda_0 c_1} + e^{-\lambda_0 c_2} = \alpha \\ c_1 e^{-\lambda_0 c_1} - c_2 e^{-\lambda_0 c_2} = 0 \end{cases}$$

la deuxième équation correspondant à la condition d'annulation de la dérivée en λ_0 de $1 - e^{-\lambda c_1} + e^{-\lambda c_2}$, comme indiqué dans la note 9.3.
Supposons que le risque soit équiréparti sur les deux extrémités. Alors on aurait $1 - e^{-\lambda_0 c_1} = e^{-\lambda_0 c_2} = \frac{\alpha}{2}$, ou $\lambda_0 c_1 = -\ln(1 - \frac{\alpha}{2})$ et $\lambda_0 c_2 = -\ln \frac{\alpha}{2}$. Ce qui donnerait par la deuxième équation :

$$\lambda_0 c_1 (1 - \tfrac{\alpha}{2}) - \lambda_0 c_2 \tfrac{\alpha}{2} = 0 \quad \Longrightarrow \quad \tfrac{\alpha}{2} \ln \tfrac{\alpha}{2} - (1 - \tfrac{\alpha}{2}) \ln(1 - \tfrac{\alpha}{2}) = 0.$$

Or on peut vérifier que la fonction $x \ln x - (1-x)\ln(1-x)$ ne s'annule pas sur $]0, \frac{1}{2}[$ et l'équirépartition n'est donc pas possible pour $\alpha \in]0, 1[$.

Exemple : $\lambda_0 = 1$, $\alpha = 0,10$ donnent la répartition $1 - e^{-\lambda_0 c_1} = 0,081$ et $e^{-\lambda_0 c_2} = 0,019$.

Exercice 9.8

On est là dans un cas particulier de l'exemple 9.8. La fonction de vraisemblance est maximisée, au dénominateur du RVG, pour $\widehat{\mu}^{MV} = \overline{x}$. Pour le numérateur on doit la maximiser, pour μ dans $[\mu_1, \mu_2]$, ce qui revient à minimiser $\mu(\mu - 2\overline{x})$. Si $\overline{x} \in [\mu_1, \mu_2]$, le minimum est atteint pour $\mu = \overline{x}$, le RVG vaut 1 et on accepte toujours.
Si $\overline{x} < \mu_1$, le minimum est atteint pour $\mu = \mu_1$ et le RVG vaut alors $\exp\{-\frac{n}{2\sigma^2}(\mu_1 - \overline{x})^2\}$. On rejette donc H_0 pour $\frac{(\overline{x} - \mu_1)^2}{\sigma^2/n} > k_1$, soit $\frac{\overline{x} - \mu_1}{\sigma/\sqrt{n}} < -k_1$. De même si $\overline{x} > \mu_2$ on rejette si $\frac{\overline{x} - \mu_2}{\sigma/\sqrt{n}} > k_2$.

Par symétrie on prend $k_1 = k_2 = k$ avec k tel que :

$$\text{pour tout } \mu \in [\mu_1, \mu_2], P_\mu\left(\frac{\overline{X}-\mu_1}{\sigma/\sqrt{n}} < -k\right) + P_\mu\left(\frac{\overline{X}-\mu_2}{\sigma/\sqrt{n}} > k\right) \leq \alpha$$

$$\text{ou } \; P_\mu\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < \frac{\mu_1-\mu}{\sigma/\sqrt{n}} - k\right) + P_\mu\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} > \frac{\mu_2-\mu}{\sigma/\sqrt{n}} + k\right) \leq \alpha.$$

On peut aisément vérifier que la somme de ces probabilités est identique pour $\mu = \mu_1$ et pour $\mu = \mu_2$. En admettant que le maximum est atteint pour ces valeurs, on doit trouver k tel que :
$\Phi(-k) + 1 - \Phi(\frac{\mu_2-\mu_1}{\sigma/\sqrt{n}} + k) = \alpha \iff \Phi(\frac{\mu_2-\mu_1}{\sigma/\sqrt{n}} + k) - \Phi(-k) = 1 - \alpha$, où Φ est la fonction de répartition de la loi $\mathcal{N}(0\,;1)$.

Application :
Il faut trouver k tel que $\Phi(\sqrt{n}+k) - \Phi(-k) = 0,95$. En prenant k =1,645 on a $\Phi(-k) = 0,05$ et pour $n \geq 2$, $\Phi(\sqrt{n}+k) \simeq 1$. On donne ci-après, pour $n = 50$, le graphe de :

$$\begin{aligned} h(\mu) &= P_\mu\left(Z < \sqrt{n}(4-\mu) - 1,645\right) + P_\mu\left(Z > \sqrt{n}(5-\mu) + 1,645\right) \\ &= \Phi(\sqrt{n}(4-\mu) - 1,645) + 1 - \Phi(\sqrt{n}(5-\mu) + 1,645). \end{aligned}$$

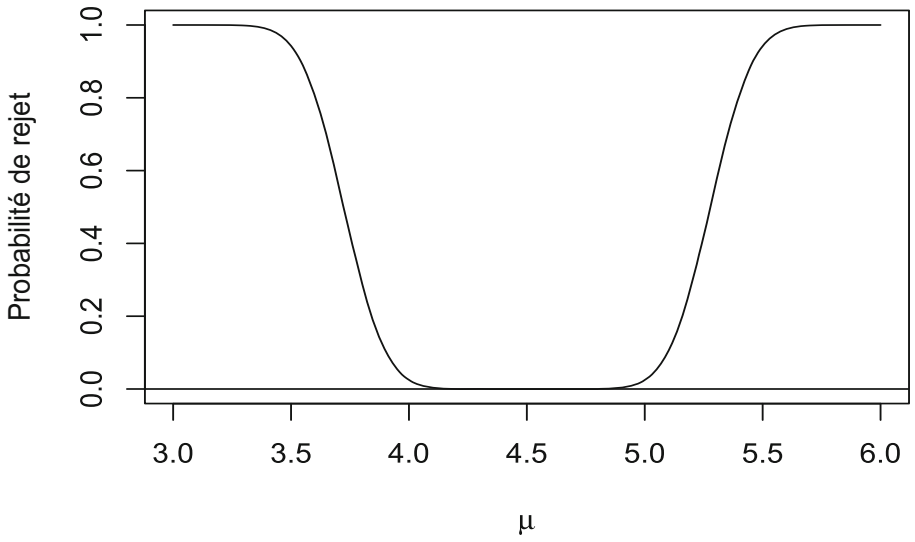


Figure 9.4 - Fonction de puissance et de risque de première espèce

Pour tester $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ appliquons les résultats précédents en prenant $\mu_1 = \mu_2 = \mu_0$.

On rejette pour $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < -k$ (cas où $\overline{x} < \mu_0$) ou $\frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > k$ (cas où $\overline{x} > \mu_0$) où k est tel que $\Phi(-k)+1-\Phi(k)=\alpha$, soit encore $1-\Phi(k)=\frac{\alpha}{2}$ ou $\Phi(k)=1-\frac{\alpha}{2}$, d'où $k=z_{1-\frac{\alpha}{2}}$. Cela est le test classique donné en section 9.7.1 et fondé sur le fait que, sous H_0, $\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \rightsquigarrow \mathcal{N}(0\,;1)$.

On a vu ci-dessus que $\Lambda_n = \exp\{-\frac{n}{2\sigma^2}(\mu_0-\overline{X})^2\}$, donc $-2\ln\Lambda_n = \left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2$. Cette variable aléatoire suit une loi $\chi^2(1)$ comme carré d'une gaussienne centrée-réduite. Le théorème asymptotique 9.2 est en fait vérifié pour tout n ici.

Exercice 9.9

On a $L(\lambda)=\lambda^n e^{-n\lambda\overline{x}}$, prenant son maximum pour $\widehat{\lambda}^{MV}=\frac{1}{\overline{x}}$, soit $L(\widehat{\lambda}^{MV}) = (\overline{x})^{-n}e^{-n}$. Pour tester $H_0:\lambda=\lambda_0$ vs. $H_1:\lambda\neq\lambda_0$ le RVG est $L(\lambda_0)/L(\widehat{\lambda}^{MV}) = (\lambda_0\overline{x})^n e^{-n(\lambda_0\overline{x}-1)}$. Appliquant le théorème 9.2, on rejettera H_0 au niveau α si $-2\ln\left((\lambda_0\overline{x})^n e^{-n(\lambda_0\overline{x}-1)}\right) > \chi^{2\,(1)}_{1-\alpha} \iff 2n[\lambda_0\overline{x}-1-\ln(\lambda_0\overline{x})] > \chi^{2\,(1)}_{1-\alpha}$.

Application :
Pour $\lambda_0=1/4$ et $n=30$, la règle de rejet avec $\alpha=0,05$ est $\frac{\overline{x}}{4}-1-\ln\frac{\overline{x}}{4} > \frac{3{,}84}{60}$. Or la fonction $g(u)=u-1-\ln u$ décroît de $+\infty$ à 0 quand x varie de 0 à 1 et croît ensuite de 0 à $+\infty$. Donc $g(u)=c$ admet deux solutions pour $c>0$. Pour $g(u)=\frac{3{,}84}{60}$, on trouve, par approximations successives, les solutions $u_1\simeq 0,68$ et $u_2\simeq 1,40$. On sera ainsi amené à rejeter H_0 si $\overline{x}$ <2,72 ou $\overline{x}$ >5,60.

Exercice 9.10

On a $L(a)=2^n a^n \sum_{i=1}^n x_i \exp\{-a\sum_{i=1}^n x_i^2\}$. Déterminons l'estimateur du MV de a :

$\ln L(a) = n\ln a - a\sum_{i=1}^n x_i^2 + \ln(2^n\sum_{i=1}^n x_i)$, $\frac{\vartheta}{\vartheta a}\ln L(a) = \frac{n}{a} - \sum_{i=1}^n x_i^2$, qui s'annule pour $\widehat{a}^{MV} = n/\sum_{i=1}^n x_i^2$.

Le RVG pour tester $H_0:a=1$ vs. $H_1:a\neq 1$ est :

$$\frac{L(1)}{L(\widehat{a}^{MV})} = \frac{2^n\sum_{i=1}^n x_i\exp\{-\sum_{i=1}^n x_i^2\}}{2^n(n/\sum_{i=1}^n x_i^2)^n\sum_{i=1}^n x_i\exp\{-n\}} = \left(\frac{\sum_{i=1}^n x_i^2}{n}\right)^n \exp\{n-\sum_{i=1}^n x_i^2\}.$$

Pour la statistique RVG, on remplace x_i par X_i dans cette expression.

Exercice 9.11

On a $L(\theta) = \left(\frac{\theta}{2}\right)^n \left(\prod_{i=1}^n \frac{2}{x_i}\right)^{\theta+1}$ si $x > a$. Déterminons l'estimateur du MV de θ :

$$\ln L(\theta) = n \ln\left(\frac{\theta}{2}\right) - (\theta+1)\sum_{i=1}^n \ln(\tfrac{x_i}{2}),\ \tfrac{\vartheta}{\vartheta\theta} \ln L(\theta) = \tfrac{n}{\theta} - \sum_{i=1}^n \ln(\tfrac{x_i}{2}),$$

qui s'annule pour $\widehat{\theta}^{MV} = n/\sum_{i=1}^n \ln\left(\frac{x_i}{2}\right)$.

Le logarithme du RVG pour tester $H_0 : \theta = 3$ vs. $H_1 : \theta \neq 3$ est :

$$\begin{aligned}\ln RVG &= n \ln\left(\frac{3}{2}\right) - 4\sum_{i=1}^n \ln(\frac{x_i}{2}) - n \ln\left(\frac{\widehat{\theta}^{MV}}{2}\right) + (\widehat{\theta}^{MV}+1)\sum_{i=1}^n \ln(\frac{x_i}{2}) \\ &= n \ln\left(\frac{3}{\widehat{\theta}^{MV}}\right) + (\widehat{\theta}^{MV}-3)\sum_{i=1}^n \ln(\frac{x_i}{2})\end{aligned}$$

et on rejette H_0 au niveau $0,05$ si $-2 \ln RVG > \chi^{2\ (1)}_{0,95} = 3,84$.

Application :
Pour $n = 30$ et $\sum_{i=1}^{30} \ln x_i = 31$ on a $\sum_{i=1}^n \ln(\frac{x_i}{2}) = 31 - 30 \ln 2$ =10,206, $\widehat{\theta}^{MV}$ =2,9396 et $-2 \ln RVG = 0,0125$. Donc on accepte H_0.
Note : H_0 spécifie une valeur proche de $\widehat{\theta}^{MV}$.

Exercice 9.12

On a $L(\lambda) = (\prod_{i=1}^n x_i!)e^{-n\lambda}\lambda^{n\overline{x}}$, $\widehat{\lambda}^{MV} = \overline{x}$ et, pour tester $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$:

$$RVG = \frac{e^{-n\lambda_0}\lambda_0^{n\overline{x}}}{e^{-n\overline{x}}\overline{x}^{n\overline{x}}},\ -2 \ln RVG = 2n[\lambda_0 - \overline{x} + \overline{x}(\ln \overline{x} - \ln \lambda_0)].$$

Étudions la fonction $g(u) = \lambda_0 - u + u(\ln u - \ln \lambda_0)$pour $u \in]0, +\infty[$. On a $g'(u) = \ln u - \ln\ \lambda_0$ et est négative pour $u < \lambda_0$, positive pour $u > \lambda_0$. Quand $u \to 0$, g$(u) \to \lambda_0$ et quand $u \to +\infty$, g$(u) \to +\infty$. *la* fonction décroît donc de λ_0 à 0 à gauche de λ_0 et croît de 0 à $+\infty$ à droite. L'inéquation $g(u) > k$ admet donc comme solutions les valeurs de u à l'extérieur de $[c_1, c_2]$ où $0 < c_1 < \lambda_0 < c_2$ si $k \in]0, \lambda_0[$ et à l'extérieur de $[0, c_2]$ où $\lambda_0 < c_2$ si $k \geq \lambda_0$.

Étant donné que le rejet de H_0 se fait (approximativement) au niveau α quand $-2 \ln RVG > \chi^{2\ (1)}_{1-\alpha}$, c'est-à-dire $g(\overline{x}) > \frac{1}{2n}\chi^{2\ (1)}_{1-\alpha}$, il y a deux cas de figure. Soit $\frac{1}{2n}\chi^{2\ (1)}_{1-\alpha} < \lambda_0$ et le rejet est fondé sur un intervalle de la forme $[c_1, c_2]$, soit $\frac{1}{2n}\chi^{2\ (1)}_{1-\alpha} \geq \lambda_0$ et il est fondé sur un intervalle de la forme $[0, c_2]$. Ce dernier

cas se présente si λ_0 est trop petit compte tenu de la taille d'échantillon et du niveau α souhaité pour pouvoir rejeter sur des valeurs de $\overline{x}$ entre 0 et λ_0. Dans le premier cas, on a la relation $-c_1 + c_1(\ln c_1 - \ln \lambda_0) = -c_2 + c_2(\ln c_2 - \ln \lambda_0)$ résultant du fait que $g(c_1) = g(c_2)$.

Application :
Pour $\lambda_0 = 5$, $n = 10$ et $\alpha = 0{,}05$, on a $\chi^{2\ (1)}_{0,95} = 3,84$ et on rejette H_0 si $g(\overline{x}) > 0,192$. Par un logiciel mathématique (ou par approximations successives) on trouve pour $g(u) = 0,192$, soit $5 - u + u(\ln u - \ln 5) = 0,192$, les solutions $c_1 = 3,68$ et $c_2 = 6,45$. On rejette donc H_0 si $\sum_{i=1}^n x_i < 36,8$ ou $\sum_{i=1}^n x_i > 64,5$. Comme $\sum_{i=1}^n x_i$ est entier, on rejettera de façon conservatrice si $\sum_{i=1}^n x_i \leq 36$ ou $\sum_{i=1}^n x_i \geq 64$.

On peut trouver le niveau exact de la règle ci-dessus dans la mesure où $\sum_{i=1}^n X_i \rightsquigarrow \mathcal{P}(n\lambda_0)$ sous H_0. On trouve, par exemple via EXCEL :

$$P(\textstyle\sum_{i=1}^n X_i \leq 36) = 0,0238 \text{ et } P(\sum_{i=1}^n X_i \geq 65) = 0,0236$$

d'où le niveau exact de 0,0474.

Pour randomiser, changeons la borne supérieure pour laquelle il faudrait atteindre une probabilité de 0,0262 pour avoir précisément $\alpha = 0,05$. Comme $P(\sum_{i=1}^n X_i \geq 64) = 0,0318$ on choisit la limite 65 avec probabilité p et 64 avec probabilité $1-p$ où $p\times 0{,}0236+(1-p)\times 0{,}0318=0{,}0262$, soit p =0,683.

Exercice 9.13

Soit à tester $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ avec un échantillon issu de la loi $\mathcal{N}(\mu\,, \sigma^2)$, σ^2 étant inconnu. D'une façon générale, on a :

$$L(\mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\}.$$

Comme $\widehat{(\mu, \sigma^2)}^{MV} = (\overline{x},\, \widetilde{s}^2)$ le dénominateur du RVG est :

$$L(\overline{x},\, \widetilde{s}^2) = \left[\frac{2\pi \sum_{i=1}^n (x_i - \overline{x})^2}{n}\right]^{-\frac{n}{2}} e^{-\frac{n}{2}}.$$

Le numérateur est la maximisation de $L(\mu, \sigma^2)$ sur $\Theta_0 = \{(\mu_0, \sigma^2), \sigma^2 > 0\}$, soit :

$$\begin{aligned}\sup_{\Theta_0} L(\mu, \sigma^2) &= \sup_{\sigma^2} \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2\} \\ &= \left[\frac{2\pi \sum_{i=1}^n (x_i - \mu_0)^2}{n}\right]^{-\frac{n}{2}} e^{-\frac{n}{2}}.\end{aligned}$$

car le sup sur σ^2 est atteint pour $\sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \mu_0)^2$. Ainsi :

$$RVG = \frac{L(\overline{x},\, \widetilde{s}^2)}{\sup\limits_{\Theta_0} L(\mu, \sigma^2)} = \left[\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]^{-\frac{n}{2}}.$$

Or $\sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2$, d'où :

$$RVG = \left[1 + \frac{n(\overline{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]^{-\frac{n}{2}} = \left[1 + \frac{(\overline{x} - \mu_0)^2}{\widetilde{s}^2}\right]^{-\frac{n}{2}}.$$

Le test du RVG consiste à rejeter H_0 si :

$$\left[1 + \frac{(\overline{x} - \mu_0)^2}{\widetilde{s}^2}\right]^{-\frac{n}{2}} < k \;\Leftrightarrow\; \frac{(\overline{x} - \mu_0)^2}{\widetilde{s}^2} > k' \;\Leftrightarrow\; \left|\frac{\overline{x} - \mu_0}{\widetilde{s}/\sqrt{n}}\right| > k''$$

ce qui est la forme du test de Student. En revanche la région de rejet, pour un α fixé, donnée par le résultat asymptotique sur le RVG ne coïncide pas avec celle du test de Student.

Exercices appliqués

Exercice 9.14

La question est de savoir si l'on peut admettre l'efficacité de la nouvelle fabrication ($\mu > 64\,000$) avec un risque d'erreur contrôlé. On doit donc tester :

$$H_0 : \mu \leq 64\,000 \text{ vs. } H_1 : \mu > 64\,000.$$

Soit $\overline{X}$ la moyenne d'un échantillon de taille 10 de pneus de nouvelle fabrication. Sous H_0 ($\mu = 64\,000$), $Z = \frac{\overline{X} - 64\,000}{8\,000/\sqrt{10}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1)$. On doit rejeter H_0 au niveau de risque $\alpha = 0,05$ si l'on observe une valeur z de Z supérieure à $z_{0,95} = 1,645$.

Dans l'expérience effectuée on a observé $\overline{x} = 67\,300$, soit $z = \frac{67\,300 - 64\,000}{8\,000/\sqrt{10}} = 1,30$ et on accepte H_0. La méthode ne peut être jugée efficace.

Calculons la puissance du test pour une valeur $\mu = 65\,000$. Dans cette alternative on a donc $\frac{\overline{X} - 65\,000}{8\,000/\sqrt{10}} \underset{approx}{\rightsquigarrow} \mathcal{N}(0\,;1)$ et, selon la règle de décision utilisée, la probabilité de rejeter H_0 est :

$$P\left(\frac{\overline{X} - 64\,000}{8\,000/\sqrt{10}} > 1,645\right) = P\left(\frac{\overline{X} - 65\,000}{8\,000/\sqrt{10}} > 1,645 - \frac{1\,000}{8\,000/\sqrt{10)}}\right) = P(Z > 1,25) = 0{,}106.$$

De façon plus générale, pour toute valeur de μ, la probabilité de rejet est :

$$h(\mu) = P\left(\frac{\overline{X} - 65\,000}{8\,000/\sqrt{10}} > 1,645 - \frac{\mu - 64\,000}{8\,000/\sqrt{10)}}\right) = 1 - \Phi\left(1,645 - \frac{\mu - 64\,000}{8\,000/\sqrt{10)}}\right)$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0\,;1)$. Pour $\mu = 67\,000$ on trouve $h(67\,000) = 0,323$, pour $\mu = 69\,000$ on trouve $h(69\,000) = 0,629$, pour $\mu = 71\,000$ on trouve $h(71\,000) = 0,869$.
Le graphe de la fonction est analogue celui de l'exemple 9.5 : la courbe part de 0,05 pour $\mu = 64\,000$ et croît comme l'indiquent les calculs ci-dessus.

Exercice 9.15

On teste $H_0 : \mu = 69\,800$ vs. $H_1 : \mu \neq 69\,800$. Sous H_0, $T = \frac{\overline{X}-5}{S/\sqrt{6}} \underset{approx}{\rightsquigarrow} t(499) \simeq \mathcal{N}(0\,;1)$. T a pris la valeur $\frac{68\,750-69\,800}{10\,350/\sqrt{500}} = -2,27$. On a $P(T < -2,27) \simeq 0{,}012$ et s'agissant d'un test bilatéral la P-valeur est 0,024 (ceci pour être en cohérence avec la notion de risque α).

Exercice 9.16

On teste $H_0 : \mu \geq 5$ vs. $H_1 : \mu < 5$, l'alternative H_1 correspondant à l'affirmation (eau non potable) dont le risque doit être contrôlé. Sous H_0, $T = \frac{\overline{X}-5}{S/\sqrt{6}} \underset{approx}{\rightsquigarrow} t(5)$ et l'on doit rejeter H_0 au niveau de risque α si T prend une valeur $t < t_\alpha^{(5)} = -t_{1-\alpha}^{(5)}$.

On a observé $\overline{x} = 4,9567$ et $s = 0,1401$, soit $t = \frac{4,9567-5}{0,1401/\sqrt{6}} = -0,757$. Pour $\alpha = 0,05$ $t_{0,05}^{(5)} = -2,105$ et il n'y a donc pas lieu de rejeter H_0. En d'autres termes on ne peut affirmer que l'eau n'est pas potable.

Exercice 9.17

L'hypothèse à tester est que le temps moyen est resté identique, soit :

$$H_0 : \mu = 42,5 \text{ vs. } H_1 : \mu \neq 42,5.$$

On rejettera H_0 au niveau 0,05 si $t = \frac{\overline{x}-42,5}{s/\sqrt{30}} \notin [-t_{0,975}^{(29)}, t_{0,975}^{(29)}]$ soit [-2,045 ;2,045].
Ici on a $t = \frac{39-42,5}{8,2/\sqrt{30}} = -2,34$ et l'on rejette H_0.
Pour $\alpha = 0,01$ on a $t_{0,995}^{(29)} = 2,756$ et il n'est pas possible de rejeter H_0 à ce niveau de risque plus faible.

Exercice 9.18

L'hypothèse à tester est $2\sigma \leq 0{,}1$, ce qui équivaut à $H_0 : \sigma^2 \leq 0{,}0025$ vs. $H_1 : \sigma^2 > 0{,}0025$. On rejettera H_0 au niveau 0,05 si $q = \frac{5s^2}{0{,}0025} > \chi^{2\ (5)}_{0{,}95} = 11,1$ et au niveau 0,01 si $q > \chi^{2\ (5)}_{0{,}99} = 15,1$ (voir section 9.7.2).

Sur la base des observations on trouve $s^2 = 0,009216$, soit $q = 18,43$ ce qui nous permet de rejeter H_0 au niveau 0,01. Note : 18,43 est le quantile 0,9975 de la loi $\chi^2(5)$ − dans EXCEL 1-LOI.KHIDEUX(18,43 ; 5) − et la P-valeur est donc 0,0025.

Exercice 9.19

On doit tester $H_0 : p \leq 0{,}04$ vs. $H_1 : p > 0{,}04$ où p est la probabilité qu'une naissance soit prématurée dans la région considérée. On suppose − ce qui est assez réaliste − que les observations effectuées sont indépendantes. Comme $np_0 = 1\,243 \times 0,04 > 5$ et aussi $n(1-p_0) > 5$ on peut appliquer l'approximation gaussienne décrite en section 9.7.5. On rejettera H_0 au niveau α si :

$$z = \frac{\widehat{p} - 0,04}{\sqrt{\frac{(0,04)(0,96)}{1243}}} > z_{1-\alpha},$$

où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0\,;1)$.

On trouve ici, avec $\widehat{p} = 72/1243 = 0{,}05792$, $z = 3,22$ ce qui correspond à une P-valeur de 0,0006 (voir la table $\mathcal{N}(0\,;1)$). On peut donc rejeter au moins au niveau 0,001. On peut donc affirmer avec un risque très faible d'erreur que la proportion (théorique) de prématurés est plus élevée dans cette région que dans le nord de l'Italie en général.

Au niveau 0,01 on est amené à rejeter si $z > z_{0,99} = 2{,}33$ ce qui équivaut à une réalisation $\widehat{p} > 0{,}0530$ de P_n, la proportion de prématurés dans un échantillon aléatoire de taille n (ici $n = 1\,243$) avec probabilité p qu'un nouveau né soit prématuré à chaque naissance. Pour calculer la puissance pour une valeur p donnée, supérieure à 0,04, il faut calculer la probabilité d'avoir une réalisation de P_n supérieure à 0,053. La fonction puissance est donc :

$$h(p) = P(P_n > 0,053) = P(Z > \frac{0,053 - p}{\sqrt{\frac{p(1-p)}{1\,243}}}) = 1 - \Phi\left(\frac{0,053 - p}{\sqrt{\frac{p(1-p)}{1\,243}}}\right)$$

pour $p > 0,04$, où $Z \rightsquigarrow \mathcal{N}(0\,;1)$ et Φ est la fonction de répartition de cette loi. Le graphe de $h(p)$ est une courbe croissante partant de 0,01 pour $p = 0,04$ (point non inclus) jusqu'à 1 quand $p = 1$. En fait $h(p)$ se rapproche très vite de 1 puisqu'elle dépasse déjà 0,999 pour $p = 0,08$. Cela est à relier à la taille d'échantillon élevée.

Exercice 9.20

On doit tester $H_0 : p \leq 0{,}04$ vs. $H_1 : p > 0{,}04$ où p est la proportion de pièces défectueuses dan le lot. On applique l'approximation gaussienne car $np_0 = 800 \times 0,04 > 5$ et $n(1-p_0) = 800 \times 0,96 > 5$.

On a trouvé $\widehat{p} = 40/800 = 0{,}05$, $z = \frac{0{,}05-0{,}04}{\sqrt{\frac{(0,04)(0,96)}{800}}} = 1,44$ ce qui correspond à une P-valeur de 0,0749 et ne suffit pas pour rejeter H_0 ne serait-ce qu'au niveau 0,05 (qui aurait nécessité une valeur de z supérieure à 1,645).

Exercice 9.21

Soit μ_1 la moyenne de taux de plomb des filles du primaire dans la ville et μ_2 celle des garçons. On se pose la question de savoir si l'on peut, à faible risque d'erreur, considérer qu'il y a une différence en moyenne, sans s'intéresser au sens de cette différence. On testera donc $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ et on rejettera H_0 au niveau α (voir section 9.7.3) si :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \notin \left[-t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}, t_{1-\frac{\alpha}{2}}^{(n_1+n_2-2)}\right] \text{ où } s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

avec $n_1 = 32$ et $n_2 = 35$. Les tailles d'échantillons sont supérieures à 30 et donc suffisantes pour appliquer la procédure, mais nous devons faire l'hypothèse supplémentaire (au sens de condition à remplir) que $\sigma_1^2 = \sigma_2^2$. Toutefois cette condition n'est pas cruciale comme on l'a indiqué en section 9.7.3 dans la mesure où les tailles d'échantillons sont proches. Qui plus est, les variances empiriques semblent indiquer que les variances théoriques sont du même ordre. On a :

$$s_p^2 = \frac{31 \times 3{,}39 + 34 \times 3{,}94}{65} = 3{,}678,\ s_p = 1{,}918 \text{ et } t = \frac{12{,}50 - 12{,}40}{1{,}918\sqrt{\frac{1}{32}+\frac{1}{35}}} = 0{,}213.$$

On ne peut rejeter H_0 au niveau 0,05 car $[-t_{0,95}^{(65)}, t_{0,95}^{(65)}] \simeq [-2{,}00;2{,}00]$. En d'autres termes, la différence observée entre les deux moyennes n'est pas significative.

Exercice 9.22

Soit μ_1 la moyenne de l'an dernier pour l'ensemble des appartements de 3 pièces en ville et μ_2 celle de cette année. Prenons l'hypothèse explicitée dans l'énoncé comme étant l'hypothèse nulle, ce qui revient à se poser la question de savoir si l'on peut, à faible risque d'erreur, considérer qu'il y a eu une augmentation en moyenne. On testera donc $H_0 : \mu_1 \geq \mu_2$ vs. $H_1 : \mu_1 < \mu_2$ et on rejettera H_0 au niveau 0,05 (voir section 9.7.3) si :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{0,95}^{(58)} = -1,67\,, \text{ où } \; s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

car une valeur de t trop négative correspond à $\overline{x}_1$ nettement inférieur à $\overline{x}_2$ ce qui va dans le sens de H_1.
Les tailles d'échantillons sont (quasi) suffisantes pour appliquer la procédure et nous supposerons pour l'heure que la condition $\sigma_1^2 = \sigma_2^2$ est remplie. On a :

$$s_p^2 = \frac{28(26)^2 + 30(28)^2}{58} = 732,\ s_p = 27,1 \text{ et } t = \frac{325 - 338}{27,1\sqrt{\frac{1}{29} + \frac{1}{31}}} = -1,86$$

ce qui nous amène à rejeter H_0 au niveau 0,05, sans plus, la P-valeur étant de 0,034 (obtenue par EXCEL : LOI.STUDENT(1,86 ; 58 ; 1) ; cette fonction donne la probabilité d'être au-delà d'une valeur − nécessairement positive − avec le paramètre 1 ; pour 2 la probabilité est doublée, ce qui correspond à une situation bilatérale) .

La procédure est applicable si les tailles d'échantillon permettent les approximations gaussiennes, ce qui est le cas ici avec des tailles proches de 30. D'autre part, la condition $\sigma_1^2 = \sigma_2^2$ n'est pas cruciale du fait que $n_1 \simeq n_2$, d'autant plus que les valeurs des variances observées sont assez voisines. Il faut aussi que les deux échantillons aient été sélectionnés indépendamment.

Exercice 9.23

Soit μ_1 la moyenne théorique (population pathologique virtuelle supposée définie) du rythme cardiaque avant traitement et μ_2 celle après traitement. On se pose la question de savoir si l'on peut, à faible risque d'erreur, considérer que le traitement est efficace, soit $\mu_2 > \mu_1$. On testera donc $H_0 : \mu_1 \geq \mu_2$ vs. $H_1 : \mu_1 < \mu_2$. Il s'agit d'un test apparié (voir en fin de section 9.7.3) car les mesures sont répétées sur le même échantillon. Ce test porte sur la

moyenne δ des différences, après moins avant par exemple, ce qui se traduit par $H_0 : \delta \leq 0$ vs. $H_1 : \delta > 0$. On rejettera H_0 au niveau α (voir les notations section 9.7.3) si $t = \frac{\overline{d}}{s_d/\sqrt{n}} > t_{1-\alpha}^{(n-1)}$, car une valeur élevée (positive) résulte d'une différence moyenne observée fortement positive, ce qui va dans le sens de H_1. Les observations après moins avant sont $4\,;\, 5{,}5\,;\, 3{,}2\,;1\,;\, -1$ et donnent une moyenne $\overline{d} = 2,54$, un écart-type $s_d = 2,561$. D'où $t = 2{,}218$.

Pour $\alpha = 0{,}05$ on a $t_{0,95}^{(4)} = 2,132$ ce qui nous amène à rejeter H_0 à ce niveau de risque. On ne peut rejeter au niveau 0,01 car la P-valeur est égale à 0,0454 (EXCEL : LOI.STUDENT(2,218 ; 4 ; 1)). Si l'on est prêt à prendre un risque d'erreur de 5% on peut admettre que le traitement est efficace.

Exercice 9.24

Comme pour l'exercice précédent, il s'agit d'un test apparié. Soit μ_1 la moyenne théorique du niveau des ventes dans l'ancien régime et μ_2 celle relative au nouveau régime (sur la population de l'ensemble des vendeurs). Si l'on veut pouvoir affirmer, avec un faible risque d'erreur, qu'il y a baisse de niveau des ventes ($\mu_2 < \mu_1$) il faut mettre cet état de fait en H_1. On testera donc H_0 : $\mu_1 \leq \mu_2$ vs. $H_1 : \mu_1 > \mu_2$, soit encore en posant $\delta = \mu_2 - \mu_1$ (nouveau moins ancien) $H_0 : \delta \geq 0$ vs. $H_1 : \delta < 0$.

On rejettera H_0 au niveau α si $t = \frac{\overline{d}}{s_d/\sqrt{n}} < -t_{1-\alpha}^{(n-1)}$, car une valeur nettement négative résulte d'une différence moyenne observée de même nature, ce qui va dans le sens de H_1. La moyenne des 16 différences observées nouveau moins ancien est trouvée égale à $\overline{d} = -2{,}315$, avec un écart-type s_d =10,675. D'où $t = -0{,}8665$. De toute évidence, on ne peut faire l'affirmation mentionnée avec une telle valeur de t ($-t_{0,95}^{(15)} = -1{,}753$ et P-valeur $\simeq 0{,}2$).

Exercice 9.25

L'idée est de voir si, sur la base d'observations, on peut considérer que le vaccin est efficace, soit, avec des notations évidentes, $p_A < p_B$. On testera donc $H_0 : p_A \geq p_B$ vs. $H_1 : p_A < p_B$. On suppose que les deux échantillons ont été sélectionnés indépendamment. On rejettera H_0 si la statistique de test prend une valeur :

$$z = \frac{\widehat{p}_A - \widehat{p}_B}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_A} + \frac{1}{n_B})}}$$

trop négative (vis-à-vis de la loi $\mathcal{N}(0\,;1)$) car cela résulterait d'une valeur $\widehat{p}_A$ nettement inférieure à $\widehat{p}_B$, ce qui va dans le sens de H_1.

Comme $n_A\widehat{p}_A(1-\widehat{p}_A) = 32 > 12$ et $n_B\widehat{p}_B(1-\widehat{p}_B) = 48 > 12$, on est en droit d'appliquer la procédure approximative décrite en section 9.7.6. On a $\widehat{p}_A = 0,20$, $\widehat{p}_B = 0,40$ et $\widehat{p} = (40{+}80)/400 = 0{,}3$. D'où $z = \frac{0,20-0,40}{\sqrt{(0,30)(0,70)(1/200+1/200)}} = -4,36$. Cette valeur est très éloignée sur la loi de Gauss $\mathcal{N}(0\,;1)$ et a une P-valeur inférieure à 0,001 car le quantile d'ordre 0,999 est la valeur 3,10 environ (voir table). Donc, sans aucun doute, le vaccin est efficace.

Exercice 9.26

Le test est $H_0 : p_A = p_B$ vs. $H_1 : p_A \neq p_B$. On rejette H_0 si la valeur de z explicitée à l'exercice précédent est trop éloignée de façon bilatérale sur la loi $\mathcal{N}(0\,;1)$. On a trouvé ici $\widehat{p}_A = 0,5204$, $\widehat{p}_B = 0,4903$ et :

$$\widehat{p} = \frac{510+505}{980+1030} = 0{,}505\,,\ z = \frac{0,5204-0,4903}{\sqrt{(0{,}505)(0{,}495)(1/980+1/1030)}} = 1{,}35.$$

Au niveau $\alpha = 0,05$, on rejette H_0 si $z \notin [-1{,}96; 1{,}96]$. Donc on ne rejette pas. La différence des estimations fournies par les deux instituts n'est pas significative.

Notons que la procédure approximative est applicable : $n_A\widehat{p}_A(1-\widehat{p}_A) > 12$ et $n_B\widehat{p}_B(1-\widehat{p}_B) > 12$.

Chapitre 10 : Tests pour variables catégorielles et tests non paramétriques

Exercice 10.1

Prenons les notations du khi-deux introduites en fin de section 10.2, soit :

- pour la loi 1 : n_{11} succès et n_{21} échecs parmi $n_{\cdot 1}$ essais, d'où $\widehat{p}_1 = \frac{n_{11}}{n_{\cdot 1}}$
- pour la loi 2 : n_{12} succès et n_{22} échecs parmi $n_{\cdot 2}$ essais, d'où $\widehat{p}_2 = \frac{n_{12}}{n_{\cdot 2}}$

et $n = n_{\cdot 1} + n_{\cdot 2}$, $\widehat{p} = \frac{n_{11}+n_{12}}{n} = \frac{n_{1\cdot}}{n}$. Montrons que $z^2 = q$.

Le numérateur de z est $\widehat{p}_1 - \widehat{p}_2 = \frac{n_{11}}{n_{\cdot 1}} - \frac{n_{12}}{n_{\cdot 2}} = \frac{n_{11}n_{22}-n_{12}n_{21}}{n_{\cdot 1}n_{\cdot 2}}$.

Au dénominateur on a $\widehat{p}(1-\widehat{p}) = \frac{n_{1\cdot}}{n}(1 - \frac{n_{1\cdot}}{n}) = \frac{n_{1\cdot}n_{2\cdot}}{n^2}$ et $\frac{1}{n_{\cdot 1}} + \frac{1}{n_{\cdot 2}} = \frac{n}{n_{\cdot 1}n_{\cdot 2}}$.
Ainsi :

$$z^2 = \frac{(\widehat{p}_1 - \widehat{p}_2)^2}{\widehat{p}(1-\widehat{p})(\frac{1}{n_{\cdot 1}} + \frac{1}{n_{\cdot 2}})} = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}.$$

Pour q, la statistique du test du khi-deux, notons que tous les quatre écarts entre observé et attendu sont égaux en valeur absolue puisque les deux tableaux ont les mêmes marges. Prenons donc la première case, on a :

$\left|n_{11} - \frac{n_{1\cdot}n_{\cdot 1}}{n}\right| = \frac{1}{n}\left|n_{11}(n_{11}+n_{12}+n_{21}+n_{22}) - (n_{11}+n_{12})(n_{11}+n_{21})\right| = \frac{1}{n}\left|n_{11}n_{22} - n_{12}n_{21}\right|$.

Donc $q = \frac{1}{n^2}(n_{11}n_{22} - n_{12}n_{21})^2(\frac{n}{n_{1\cdot}n_{\cdot 1}} + \frac{n}{n_{1\cdot}n_{\cdot 2}} + \frac{n}{n_{2\cdot}n_{\cdot 1}} + \frac{n}{n_{2\cdot}n_{\cdot 2}})$

Comme le dernier facteur est égal à $\frac{n_{2\cdot}n_{\cdot 2}+n_{2\cdot}n_{\cdot 1}+n_{1\cdot}n_{\cdot 2}+n_{1\cdot}n_{\cdot 1}}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}} = \frac{n^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$, la propriété est démontrée.

La statistique de test Z suivant la loi $\mathcal{N}(0\,;1)$, $Z^2 \rightsquigarrow \chi^2(1)$ et les deux tests sont donc équivalents (voir avec plus de précision la section10.1.4). Notons toutefois que ceci n'est vrai que pour le test avec l'alternative bilatérale $p_1 \neq p_2$.

Exercice 10.2

Rappelons que les deux variables catégorielles $\mathcal{X}$ et $\mathcal{Y}$ sont dites indépendantes si tout événement sur l'une est indépendant de tout événement sur l'autre, un événement étant un sous-ensemble de catégories.
La condition indiquée dans l'énoncé est évidemment nécessaire puisque chaque catégorie constitue un événement en soi. Il reste à montrer qu'elle est suffisante.

D'une façon générale, si A est indépendant de B_1 d'une part et de B_2 d'autre part, B_1 et B_2 étant disjoints, alors A est indépendant de $B_1 \cup B_2$. Clairement, comme $A \cap B_1$ et $A \cap B_2$ sont disjoints on a :

$$\begin{aligned} P\left(A \cap (B_1 \cup B_2)\right) &= P((A \cap B_1) \cup (A \cap B_2)) = P(A \cap B_1) + P(A \cap B_2) \\ &= P(A)P(B_1) + P(A)P(B_2) = P(A)\left[P(B_1) + P(B_2)\right] \\ &= P(A)P(B_1 \cup B_2). \end{aligned}$$

Dans le contexte de l'énoncé, cela implique que $\{i_1\}$ étant indépendant de $\{j_1\}$ et de $\{j_2\}$ alors $\{i_1\}$ est indépendant de $\{j_1, j_2\}$ et, de proche en proche, $\{i_1\}$ est indépendant de $\{j_1, j_2, \cdots, j_k\}$ quel que soit le sous-ensemble de catégories de $\mathcal{Y}$. Par symétrie $\{i_1, i_2\}$ est indépendant de $\{j_1, j_2, \cdots, j_k\}$ et, de proche en proche, $\{i_1, i_2, \cdots, i_l\}$ est indépendant de $\{j_1, j_2, \cdots, j_k\}$.

Exercice 10.3

Soient les événements $A = (N_{11} = n_{11})$, $B = (N_{1.} = n_{1.}, N_{2.} = n_{2.}, N_{.1} = n_{.1}, N_{.2} = n_{.2})$, on souhaite évaluer $P(A|B) = P(A \cap B)/P(B)$.

Pour l'événement $A \cap B$, $N_{11} = n_{11}$ et les marges sont données, alors le tableau 2×2 est entièrement déterminé. Ainsi cet événement est identique à $(N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22})$ à condition de prendre $n_{12} = n_{1.} - n_{11}$, $n_{21} = n_{.1} - n_{11}$ et $n_{22} = n - (n_{11} + n_{12} + n_{21}) = n + n_{11} - n_{1.} - n_{.1}$. La loi conjointe de $(N_{11}, N_{12}, N_{21}, N_{22})$ est une loi multinomiale de paramètres $p_{1.}p_{.1},\ p_{1.}p_{.2},\ p_{2.}p_{.1},\ p_{2.}p_{.2}$ sous l'indépendance d'où (voir section 4.1.6) :

$$\begin{aligned} P(A \cap B) &= P(N_{11} = n_{11}, N_{12} = n_{12}, N_{21} = n_{21}, N_{22} = n_{22}) \\ &= \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!}(p_{1.}p_{.1})^{n_{11}}(p_{1.}p_{.2})^{n_{12}}(p_{2.}p_{.1})^{n_{21}}(p_{2.}p_{.2})^{n_{22}}. \end{aligned}$$

L'événement B est identique à $(N_{1.} = n_{1.}, N_{.1} = n_{.1})$ car $N_{2.}$ et $N_{.2}$ sont alors déterminés. On a $N_{1.} \rightsquigarrow B(n, p_{1.})$, $N_{.1} \rightsquigarrow B(n, p_{.1})$, avec $N_{1.}$ et $N_{.1}$ indépendants sous l'hypothèse d'indépendance des deux variables catégorielles. Donc :

$$P(B) = P(N_{1.} = n_{1.}, N_{.1} = n_{.1}) = \frac{n!}{n_{1.}!n_{2.}!}p_{1.}^{n_{1.}}p_{2.}^{n_{2.}}\frac{n!}{n_{.1}!n_{.2}!}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}\,.$$

En notant que $(p_{1.}p_{.1})^{n_{11}}(p_{1.}p_{.2})^{n_{12}}(p_{2.}p_{.1})^{n_{21}}(p_{2.}p_{.2})^{n_{22}} = p_{1.}^{n_{1.}}p_{2.}^{n_{2.}}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}$ et en faisant le rapport $P(A \cap B)/P(B)$ on obtient immédiatement le résultat à démontrer.

Exercice 10.4

1. D'une façon générale, on a la fonction de vraisemblance relative à la loi multinomiale avec les 4 catégories du tableau, soit :

$$L(p_{11}, p_{12}, p_{21}, p_{22}) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} p_{11}^{n_{11}} p_{12}^{n_{12}} p_{21}^{n_{21}} p_{22}^{n_{22}}.$$

Sous H_0 on obtient, en intégrant les contraintes $p_{21} = p_{12}$, $p_{22} = 1 - p_{11} - 2p_{12}$ (et sachant que $n_{22} = n - n_{11} - n_{12} - n_{21}$) :

$$L(p_{11}, p_{12}) = \frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} {p_{11}}^{n_{11}} p_{12}^{n_{12}+n_{21}} (1 - p_{11} - 2p_{12})^{n_{22}}.$$

2. Maximisons :

$$\ln L(p_{11}, p_{12}) = n_{11} \ln p_{11} + (n_{12}+n_{21}) \ln p_{12} + n_{22} \ln(1-p_{11}-2p_{12}) + \text{Constante}$$

en annulant les deux dérivées partielles par rapport à p_{11} et p_{12} :

$$\begin{cases} \frac{\delta}{\delta p_{11}} L(p_{11}, p_{12}) = \frac{n_{11}}{p_{11}} - \frac{n_{22}}{1-p_{11}-2p_{12}} \\ \frac{\delta}{\delta p_{12}} L(p_{11}, p_{12}) = \frac{n_{12}+n_{21}}{p_{12}} - \frac{2n_{22}}{1-p_{11}-2p_{12}} \end{cases}.$$

En annulant on obtient :

$$\frac{n_{11}}{p_{11}} = \frac{n_{22}}{1 - p_{11} - 2p_{12}} = \frac{n_{12} + n_{21}}{2p_{12}} = \frac{n}{1}$$

d'où les estimations du MV sous H_0 : $\widehat{p}_{11} = \frac{n_{11}}{n}$, $\widehat{p}_{12} = \frac{n_{12}+n_{21}}{2n} = \widehat{p}_{21}, \widehat{p}_{22} = \frac{n_{22}}{n}$.

3. Les fréquences attendues sous H_0 sont respectivement :

$$n\widehat{p}_{11} = n_{11},\ n\widehat{p}_{12} = \tfrac{n_{12}+n_{21}}{2} = n\widehat{p}_{12},\ n\widehat{p}_{22} = n_{22},$$

d'où :

$$\begin{aligned} q &= \frac{0}{n_{11}} + \frac{(n_{12} - \frac{n_{12}+n_{21}}{2})^2}{\frac{n_{12}+n_{21}}{2}} + \frac{(n_{21} - \frac{n_{12}+n_{21}}{2})^2}{\frac{n_{12}+n_{21}}{2}} + \frac{0}{n_{22}} \\ &= \frac{2}{n_{12}+n_{21}} \left[\frac{(n_{12}-n_{21})^2}{4} + \frac{(n_{21}-n_{11})^2}{4} \right] = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}. \end{aligned}$$

4. Le nombre de catégories est 4 et il y a 2 paramètres à estimer sous H_0, p_{11} et p_{12}, donc selon le théorème 10.1 (Cramer) les degrés de liberté sont $4-1-2 = 1$. Note : Pour le test du RVG, dont on sait qu'il est asymptotiquement équivalent au test du khi-deux, on aurait un seul paramètre spécifé par H_0 et on retrouve le même degré de liberté.

Exercice 10.5

Raisonnons sur des réalisations en supposant l'absence de valeurs identiques. Si $R_{(1)}$ prend la valeur r_1, cela signifie qu'il y a $r_1 - 1$ valeurs des Y_j qui précèdent $X_{(1)}$. Si $R_{(2)}$ prend la valeur r_2, il y a $r_2 - 2$ valeurs des Y_i qui précèdent $X_{(2)}$ et, plus généralement, si $R_{(i)}$ prend la valeur r_i, il y a $r_i - i$ valeurs des Y_j qui précèdent $X_{(i)}$.

En section 10.5.4, nous avons défini la statistique U de Mann-Whitney en déterminant pour chaque Y_j le nombre de X_i qui lui sont de valeur supérieure puis en totalisant sur tous les Y_j. De façon équivalente, on peut comptabiliser pour chaque $X_{(i)}$ le nombre de Y_j qui lui sont de valeur inférieure et totaliser. Dans les notations ci-dessus la réalisation u de U vaut :

$$u = r_1 - 1 + r_2 - 2 + \cdots + r_{n_1} - n_1 = \sum_{i=1}^{n_1} r_i - \frac{1}{2} n_1(n_1 + 1)$$

où $\sum_{i=1}^{n_1} r_i$ est la réalisation de la statistique T_{n_1} de Wilcoxon fondée sur les X_i.
On a donc bien la relation $U = T_{n_1} - \frac{1}{2} n_1(n_1 + 1)$.

Exercice 10.6

La corrélation de Spearman est la corrélation linéaire usuelle (voir section 3.4) entre la série des rangs R_i des X_i et la série des rangs S_i des Y_i. Chaque série étant constituée des nombres 1 à n, sa somme vaut $\frac{1}{2} n(n + 1)$, sa moyenne $\frac{1}{2}(n + 1)$, la somme des carrés $\frac{1}{6} n(n + 1)(2n + 1)$. On a :

$$R_s = \frac{\sum_{i=1}^{n} (R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \overline{R})^2 \sum_{i=1}^{n} (S_i - \overline{S})^2}} = \frac{\sum_{i=1}^{n} R_i S_i - n\overline{RS}}{\sqrt{(\sum_{i=1}^{n} R_i^2 - n\overline{R}^2)(\sum_{i=1}^{n} S_i^2 - n\overline{S}^2)}}.$$

Or $\sum_{i=1}^{n} R_i^2 - n\overline{R}^2 = \frac{1}{6} n(n+1)(2n+1) - \frac{1}{4} n(n+1)^2 = \frac{1}{12} n(n^2 - 1)$ et de même pour $\sum_{i=1}^{n} S_i^2 - n\overline{S}^2$.
Finalement en substituant dans l'expression de R_s ci-dessus on obtient immédiatement le résultat recherché.

Exercices appliqués

Les exercices qui suivent concernent des tests de type khi-deux sur des variables catégorielles. Rapportons-nous encore à EXCEL, disponible pour tous. Ces tests peuvent être effectués par la fonction TEST.KHIDEUX(plage1 ;plage2) où la plage 1 contient les effectifs observés et est soit un vecteur, soit un tableau, et plage 2 les effectifs attendus correspondants. C'est à l'utilisateur de calculer les effectifs attendus, le tableur se prêtant aisément à leur calcul. La fonction renvoie directement la P-valeur sans indiquer la valeur q prise par la statistique Q (appelée usuellement le khi-deux). On peut toutefois récupérer q par la fonction KHIDEUX.INVERSE(P-valeur ; d.l.l.).

Pour un tableau $I \times J$, que ce soit lors d'un test de comparaison de lois multinomiales ou d'indépendance de deux variables catégorielles, les degrés de liberté (d.d.l.) sont $(I-1)(J-1)$, sauf situations très particulières et rares, et le calcul de q est identique dans les deux types de situations.

Pour un vecteur à I composantes du test d'adéquation à une loi multinomiale, EXCEL applique des degrés de liberté égaux à $I-1$, mais ceci peut être incorrect lorsqu'on a dû estimer un certain nombre de paramètres de la loi de référence, car il faut encore soustraire ce nombre. On peut toutefois revenir à la valeur de q par la fonction KHIDEUX.INVERSE avec $I-1$ degrés de liberté et lire la bonne P-valeur via la fonction LOI.KHIDEUX avec le nombre de degrés de liberté correct. Nous verrons des exemples de cela.

Pour les tableaux les calculs des effectifs attendus sont simples et répétitifs. Nous donnerons ces calculs sur un premier exemple et les omettrons par la suite.

Exercice 10.7

A priori on peut penser que l'on a sélectionné les 202 individus au hasard et donc que toutes les marges sont aléatoires. Il s'agit alors de tester l'indépendance entre sexe et niveau de gêne. Toutefois, si la sélection a été faite avec des quotas par sexe, les effectifs 120 et 82 ont été fixés par le plan de sondage et il s'agit *stricto sensu* du test de comparaison de la distribution de la variable gêne entre femmes et hommes, ce qui ne change rien à la procédure du test. Le tableau des effectifs attendus sous H_0 (indépendance gêne/sexe ou identité des distributions par sexe le cas échéant) est le suivant :

Gêne \ Sexe	Femmes	Hommes	Tous
Aucune	65,3	44,7	110
Faible	30,9	21,1	52
Moyenne	14,9	10,1	25
Forte	8,9	6,1	15
	120	82	202

Rappelons qu'il s'obtient par le produit des marges, par exemple $65{,}3 = \frac{110\times 120}{202}$.
La valeur prise par la statistique du khi-deux est :

$$q = \frac{(75-65,3)^2}{65,3} + \cdots + \frac{(12-6,1)^2}{6,1} = 16{,}7.$$

Sous l'hypothèse H_0, la statistique de test Q suit asymptotiquement une loi $\chi^2(3)$. Comme aucun effectif attendu n'est inférieur à 5, l'approximation asymptotique est satisfaisante. On rejette ici au niveau $\alpha = 0,05$ car $q > \chi^{2\ (3)}_{0,95}$ =7,815 (voir table). On rejette même au niveau de risque très faible de 0,001 puisque $\chi^{2\ (3)}_{0,999}$ =16,26. De fait, la P-valeur fournie par EXCEL est de 0,0008.
On peut affirmer avec un très faible risque d'erreur qu'il y a un effet sexe pour la gêne. La comparaison des tableaux observé et attendu (sous H_0) montre que les femmes se déclarent moins sensibles à la gêne que les hommes.

Exercice 10.8

Il s'agit d'un test d'adéquation à un certain modèle de loi multinomiale. Les 4 catégories correspondent au nombre de garçons dans une famille de 3 enfants. Sous l'hypothèse H_0 que le sexe d'un nouveau né est indépendant au cours des naissances successives et que la probabilité p d'avoir un garçon à chaque naissance reste constante, le nombre de garçons X pour une famille de 3 enfants suit une loi $B(3,p)$. Pour les valeurs $x = 0,1,2,3$ les probabilités sont respectivement $(1-p)^3, p(1-p)^2, p^2(1-p), p^3$. Mais p est inconnu et doit être estimé par l'estimateur du MV qui est la proportion $\widehat{p}$ de garçons dans l'échantillon, soit :

$$\widehat{p} = \tfrac{1}{3\times 825}(0\times 71 + 1\times 297 + 2\times 336 + 3\times 121) = 0{,}5382.$$

Le tableau ci-après donne les effectifs observés en regard des effectifs attendus.

Nombre de garçons	x	0	1	2	3
Effectifs observés	n_i	71	297	336	121
Prob. théoriques estimées	$\widehat{p}_i$	$0,0985$	$0,3443$	$0,4013$	$0,1559$
Effectifs attendus	$n\widehat{p}_i$	$81,26$	$284,08$	$331,06$	$128,60$

d'où $q = \frac{(71-81,26)^2}{81,26} + \frac{(297-284,08)^2}{284,08} + \frac{(336-331,06)^2}{331,06} + \frac{(121-128,60)^2}{128,60} = 2,41$.

Sous H_0, Q suit approximativement une loi du khi-deux avec 4-1-1=2 d.d.l. du fait qu'il a fallu estimer un paramètre (les effectifs attendus étant élevés l'approximation doit être très satisfaisante). La valeur de q trouvée correspond à une P-valeur de 0,30 et il n'y a pas lieu de rejeter H_0.

Ici nous rencontrons une situation où EXCEL donne une valeur erronée de la P-valeur. En effet la fonction TEST.KHIDEUX prend un nombre de d.d.l. égal à 3, ignorant qu'un paramètre a dû être estimé. La P-valeur donnée est en fait 0,4925 laquelle, par KHIDEUX.INVERSE(0,4925 ; 3), redonne la bonne valeur de $q = 2,41$. Il faut ensuite exécuter LOI.KHIDEUX(2,41 ; 2) pour trouver la bonne P-valeur.

Exercice 10.9

Cet exercice se fait sur le même modèle que le précédent. La distribution multinomiale est donnée avec 5 catégories. Il est nécessaire d'estimer le paramètre λ de la loi de Poisson pour estimer les probabilités des catégories :

$\widehat{\lambda}^{MV} = \overline{x} = \frac{1}{50}(21 \times 0 + 18 \times 1 + 7 \times 2 + 3 \times 3 + 1 \times 4) = 0,90$.

Les probabilités théoriques du modèle sous H_0 sont calculées selon $\frac{e^{-0,90}(0,90)^x}{x!}$.
D'où le tableau :

Nombre d'accidents	x	0	1	2	3	4
Effectifs observés	n_i	21	18	7	3	1
Prob. théoriques estimées	$\widehat{p}_i$	$0,4066$	$0,3659$	$0,1647$	$0,0494$	$0,0135$
Effectifs attendus	$n\widehat{p}_i$	$20,33$	$18,30$	$8,235$	$2,470$	$0,675$

On remarque que pour x=3 et $x = 4$, les effectifs attendus sont inférieurs à 5. Pour éviter cela, il faut regrouper les 3 dernières catégories. Cette nouvelle catégorie a pour effectifs observés 11 et attendus 11,38.
La valeur prise par la statistique du khi-deux est :

$$q = \frac{(21-20,33)^2}{20,33} + \frac{(18-18,30)^2}{18,30} + \frac{(11-11,38)^2}{11,38} = 0,043.$$

Sous H_0, Q suit approximativement une loi du khi-deux avec $3-1-1=1$ d.d.l. et on rejette H_0 au niveau 0,05 si $q > \chi^{2\,(1)}_{0,95} = 3,84$. On accepte donc H_0. Note : La P-valeur est 0,836 (voir exercice précédent pour son calcul).

Exercice 10.10

On se rapportera à l'exercice 10.7 pour la procédure qui est ici analogue. On trouve (via EXCEL, par exemple) une valeur de q égale à 2,78 correspondant à une P-valeur de 0,43 sur la loi $\chi^2(3)$. Il n'a donc pas lieu de rejeter l'hypothèse d'indépendance.
Sur la base de cette enquête on ne peut considérer qu'il y ait un mode de logement différent entre étudiantes et étudiants.

Exercice 10.11

On se rapportera également à l'exercice 10.7 pour la procédure.
On est ici dans une situation de comparaison de 4 distributions multinomiales car on peut considérer qu'on a constitué 4 échantillons de tailles définies, respectivement 106, 174, 134 et 76. On trouve (via EXCEL, par exemple) une valeur de q égale à 32,87 correspondant à une P-valeur inférieure à 10^{-3} sur la loi $\chi^2(6)$. On peut considérer avec un risque quasi nul de se tromper qu'il y a une relation entre l'âge et le type d'opérateur choisi.

Exercice 10.12

On se rapportera à l'exercice 10.7 pour la procédure.
On est ici dans une situation de comparaison de 2 distributions multinomiales. On doit rejeter au niveau 0,05 pour $q > \chi^{2\,(3)}_{0,95} = 7,81$. On trouve (via EXCEL, par exemple) une valeur de q égale à 2,81 (P-valeur=0,42). Il n'y a donc pas lieu de considérer, sur la base de cette enquête, que la répartition des revenus soit différente entre les deux pays.
Note : Un des effectifs attendus vaut 4,44 et est donc inférieur à 5. On a vu toutefois en fin de section 10.2 que, pour des tableaux autres que 2×2, la condition n'était pas cruciale. De plus, ici, on est suffisamment loin de rejeter pour se soucier du niveau d'approximation.

Exercice 10.13

Le tableau ci-après donne les éléments utiles pour effectuer le test d'adéquation du khi-deux (colonnes 2 à 4) et celui de Kolmogorv-Smirnov (colonnes 5 à 8).

Mesure	Effectif Observé	Proba. théorique	Effectif attendu	Eff. Obs. cumulé	Fn. Rép. empirique	Fn. Rép. théorique	Différence test K-S
33	3	0,0010	5,7	3	0,0005	0,0010	-0,0005
34	18	0,0037	21,2	21	0,0037	0,0047	-0,0010
35	81	0,0126	72,3	102	0,0178	0,0173	0,0005
36	185	0,0348	199,7	287	0,0500	0,0521	-0,0021
37	420	0,0758	434,9	707	0,1232	0,1279	-0,0047
38	749	0,1303	747,7	1456	0,2537	0,2582	-0,0045
39	1073	0,1779	1020,8	2529	0,4407	0,4361	0,0046
40	1079	0,1920	1101,7	3608	0,6288	0,6281	0,0007
41	934	0,1643	942,8	4542	0,7916	0,7924	-0,0008
42	658	0,1112	638,1	5200	0,9062	0,9036	0,0026
43	370	0,0597	342,6	5570	0,9707	0,9633	**0,0074**
44	92	0,0253	145,2	5662	0,9868	0,9886	-0,0018
45	50	0,0086	49,3	5712	0,9955	0,9972	-0,0017
46	21	0,0022	12,6	5733	0,9991	0,9994	-0,0003
47	4	0,0005	2,9	5737	0,9998	0,9999	-0,0001
48	1	0,0001	0,6	5738	1,0000	1,0000	0,0000

La moyenne observée est 39,83 et l'écart-type 2,050. Les probabilités théoriques sous H_0 doivent être estimées sur la loi de $X \rightsquigarrow \mathcal{N}(39{,}83\,;\,(2{,}050)^2)$.
Pour la première ligne on calcule $P(X < 33,5) = 0{,}0010$, pour la deuxième $P(33,5 < X < 34,5) = 0{,}0037$, etc.

Pour le test du khi-deux il faut regrouper les 3 dernières lignes pour avoir un effectif attendu supérieur à 5, se ramenant ainsi à 14 classes. Sous H_0 (la répartition est gaussienne) la statistique de test Q a une loi asymptotique du khi-deux avec $14-1-2 = 11$ d.d.l. car il a fallu estimer moyenne et variance de la loi. On trouve une valeur $q = 36{,}1$ pour cette statistique ce qui correspond à une P-valeur de 0,00016. On peut rejeter l'hypothèse gaussienne avec un risque d'erreur infime. Ce résultat ne doit pas surprendre car la taille de l'échantillon étant grande, la puissance du test est élevée et un modèle plus adaptatif doit être recherché.

Pour le test de Kolmogorov-Smirnov on obtient une différence maximale entre fonctions de répartition empirique et théorique $d_n = 0{,}0074$. Vu la taille d'échantillon on a, pour la statistique de test D_n, une approximation de la valeur critique au niveau 0,05 qui est $1{,}36/\sqrt{n} = 0{,}018$. On voit que ce test ne rejette pas l'hypothèse gaussienne. En réalité, il est inadapté car beaucoup trop conservateur et donc peu puissant, d'une part du fait des gros effectifs sur les

classes qui ne permettent pas de suivre l'évolution détaillée de la fonction de répartition empirique, d'autre part du fait que les écarts sont sous-évalués en prenant comme référence la loi de Gauss dont la moyenne et la variance sont estimées sur les données.

Exercice 10.14

En fusionnant les deux échantillons, il y a 23 observations en-dessous de 3 et 24 au-dessus de 3. La médiane étant la valeur 3 on est contraint d'évacuer les 13 observations situées sur 3 qui sont inclassables. On raisonne donc sur une taille globale réduite à 47 pour laquelle il y a 23 valeurs sous la médiane.

Sous H_0 (pas de différence d'attitude) la loi du nombre $\widetilde{N}_1$ d'observations sous la médiane pour le premier échantillon (ménages avec enfant(s)) de taille 25, est la loi $\mathcal{H}(47\,;23\,;25)$ qui peut être approchée par une loi de Gauss de moyenne et variance :

$$\mu = \frac{25 \times 46}{2 \times 47} = 12,234; \quad \sigma^2 = \frac{25 \times 22 \times 48}{4 \times (47)^2} = 2,988\,.$$

Donc on rejettera au niveau 0,05 si $\widetilde{N}_1$ sort de l'intervalle $12{,}234 \pm 1{,}96\sqrt{2,988}$, soit [8,85 ;15,62]. Comme $\widetilde{N}_1$ a pris la valeur 15, on ne peut rejeter l'hypothèse d'absence de différence d'attitude.
La probabilité de dépasser 14,5 sur la loi de Gauss étant égale à 0,095, la P-valeur est approximativement 0,19. Sur la loi $\mathcal{H}(47\,;\,23\,;\,25)$, la probabilité $P(\widetilde{N}_1 \geq 15)$ vaut 0,092 (calculable par EXCEL) et la P-valeur exacte est donc 0,184 ce qui montre que l'approximation est très bonne.

Note : Rappelons que le test de la médiane est peu puissant d'une façon générale et qu'il l'est particulièrement ici au vu de l'étroitesse de l'échelle et de la taille des échantillons.

Chapitre 11 : Régression linéaire

Exercice 11.1

Pour $p = 2$ la formule générale de la densité du vecteur (X, Y) gaussien est :

$$f_{X,Y}(x,y) = \frac{1}{2\pi(\det \mathbf{\Sigma})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu})^t\ \mathbf{\Sigma}^{-1}\ (\mathbf{u}-\boldsymbol{\mu})\right\}.$$

où $\mathbf{u} = \binom{x}{y}$ et $\boldsymbol{\mu} = \binom{\mu_X}{\mu_Y}$.
La matrice des variances-covariances est :

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

d'où $\det\Sigma = \sigma_X^2\sigma_Y^2(1-\rho^2)$ et :

$$\Sigma^{-1} = \frac{1}{\sigma_X^2\sigma_Y^2(1-\rho^2)} \begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}.$$

L'expression de $f_{X,Y}(x,y)$ est finalement (pour plus de détails sur le calcul matriciel voir l'exercice 3.8) :

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{(1-\rho^2)}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]\right\}.$$

Par ailleurs $f_X(x) = \frac{1}{\sigma_X}\exp\left\{-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right\}$ et donc $f_{Y|X=x}(y)$ a pour constante devant l'exponentielle $\frac{1}{\sqrt{2\pi}\sigma_Y\sqrt{(1-\rho^2)}}$ et pour l'exponentielle :

$$\begin{aligned}
&-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - (1-\rho^2)\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right] \\
&= -\frac{1}{2(1-\rho^2)}\left[\rho^2\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right] \\
&= -\frac{1}{2\sigma_Y^2(1-\rho^2)}\left\{y - \left[\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]\right\}^2
\end{aligned}$$

ce qui met en évidence une loi de Gauss de moyenne $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$ et de variance $\sigma_Y^2(1-\rho^2)$.
Cela montre que dans le modèle de régression de Y sur X la droite de régression a pour pente $\rho\frac{\sigma_Y}{\sigma_X}$ et passe par le point de coordonnées (μ_X, μ_Y). De plus, la variance de Y est réduite du facteur $1-\rho^2$, c'est-à-dire d'autant plus que Y est liée à X.

Exercice 11.2

On a :

$$E(\varphi(X)) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} y \frac{f_{X,Y}(x,y)}{f_X(x)} dy \right] f_X(x) dx = \iint_{\mathbb{R}^2} y f_{X,Y}(x,y) dx dy = \mu_Y$$

qui est donc égal à $\int_{\mathbb{R}} (\beta_0 + \beta_1 x)\, f_X(x) dx = \beta_0 + \beta_1 \mu_X$.
De la même façon :

$$E(X\varphi(X)) = \iint_{\mathbb{R}^2} xy f_{X,Y}(x,y) dx dy = E(XY)$$

qui est donc égal à $\int_{\mathbb{R}} x(\beta_0 + \beta_1 x)\, f_X(x) dx = \beta_0 \mu_X + \beta_1 E(X^2)$. D'où à résoudre en (β_0, β_1) :

$$\left\{ \begin{array}{c} \beta_0 + \beta_1 \mu_X = \mu_Y \\ \beta_0 \mu_X + \beta_1 E(X^2) = E(XY) \end{array} \right. .$$

En multipliant la première équation par $-\mu_X$ et lui ajoutant la suivante, on a :

$$\beta_1 [E(X^2) - \mu_X^2] = E(XY) - \mu_X \mu_Y \quad \text{soit} \quad \beta_1 = \frac{cov(X,Y)}{\sigma_X^2} = \rho \frac{\sigma_Y}{\sigma_X}$$

et en remplaçant dans la première :
$\beta_0 = \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} \mu_X$ et ainsi $\beta_0 + \beta_1 x = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$.

Pour la variance conditionnelle :

$$\begin{aligned} E(\psi(X)) &= \int_{\mathbb{R}} V(Y|X=x) f_X(x) dx \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \left\{ y^2 - \left[\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \right]^2 \right\} \frac{f_{X,Y}(x,y)}{f_X(x)} dy \right] f_X(x) dx \\ &= \iint_{\mathbb{R}^2} \left\{ y^2 - \left[\mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \right]^2 \right\} f_{X,Y}(x,y) dx dy \\ &= E(Y^2) - \left[\mu_Y^2 + 2\mu_Y \rho \frac{\sigma_Y}{\sigma_X} E(X - \mu_X) + \rho^2 \frac{\sigma_Y^2}{\sigma_X^2} V(X) \right] \\ &= E(Y^2) - \mu_Y^2 - \rho^2 \sigma_Y^2 = \sigma_Y^2 (1 - \rho^2). \end{aligned}$$

Mais d'autre part, comme $V(Y|X=x)$ ne dépend pas de x on peut dire, à partir de la première équation ci-dessus, que $E(\psi(X)) = V(Y|X=x)$, ce qui démontre le résultat donné.

Ainsi on peut conclure que, pour tout modèle à espérance conditionnelle linéaire et à variance conditionnelle constante, les expressions de ces dernières restent identiques à celles trouvées dans le cas gaussien.

Exercice 11.3

Tout d'abord, reprenons les formules de la section 11.2.2 donnant $\widehat{\beta}_0$ et $\widehat{\beta}_1$ en raisonnant sur les réalisations y_i des Y_i :

$$\begin{cases} \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \\ \widehat{\beta}_1 = \dfrac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \end{cases}$$

dont on déduit que $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i = \overline{y} + \widehat{\beta}_1 (x_i - \overline{x})$.

De $y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y})$ découle :

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2 + \sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 + 2 \sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}).$$

Il suffit de montrer que le dernier terme à droite est nul pour avoir la décomposition utilisée en section 11.2.4. On a :

$$\sum_{i=1}^n (y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = \sum_{i=1}^n \widehat{y}_i (y_i - \widehat{y}_i) - \overline{y} \sum_{i=1}^n (y_i - \widehat{y}_i),$$

or, en substituant l'expression donnée plus haut pour $\widehat{y}_i$:

$\sum_{i=1}^n (y_i - \widehat{y}_i) = \sum_{i=1}^n \left[(y_i - \overline{y}) - \widehat{\beta}_1 (x_i - \overline{x}) \right] = 0$

car $\sum_{i=1}^n (y_i - \overline{y})$ et $\sum_{i=1}^n (x_i - \overline{x})$ sont nuls. Il reste donc le terme :

$$\begin{aligned} \sum_{i=1}^n \widehat{y}_i (y_i - \widehat{y}_i) &= \sum_{i=1}^n \left[\overline{y} + \widehat{\beta}_1 (x_i - \overline{x}) \right] (y_i - \widehat{y}_i) \\ &= \overline{y} \sum_{i=1}^n (y_i - \widehat{y}_i) + \widehat{\beta}_1 \sum_{i=1}^n (x_i - \overline{x})(y_i - \widehat{y}_i) \\ &= 0 + \widehat{\beta}_1 \left[\sum_{i=11}^n (x_i - \overline{x}) \left[y_i - \overline{y} - \widehat{\beta}_1 (x_i - \overline{x}) \right] \right] \\ &= \widehat{\beta}_1 \left[\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) - \widehat{\beta}_1 \sum_{i=11}^n (x_i - \overline{x})^2 \right] \end{aligned}$$

et ce terme est bien nul en vertu de l'équation qui donne $\widehat{\beta}_1$.

Exercice 11.4

La prévision ponctuelle de Y_0 est $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$. Comme $E(Y_0) = \beta_0 + \beta_1 x_0$ et que $\widehat{\beta}_0$, $\widehat{\beta}_1$ sont sans biais, $E(Y_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)) = 0$.
En utilisant les résultats de la section 11.2.2 sur les espérances, variances et covariance de $\widehat{\beta}_0$ et $\widehat{\beta}_1$ on a :

$$\begin{aligned} V(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) &= V(\widehat{\beta}_0) + x_0^2 V(\widehat{\beta}_1) + 2x_0\, cov(\widehat{\beta}_0, \widehat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) + x_0^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} - 2x_0 \frac{\overline{x}\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right), \end{aligned}$$

et comme Y_0 est de variance σ^2 et indépendante de $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$:

$$V(Y_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right).$$

et $Y_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$ est gaussien.

Selon la proposition 11.1 S^2 est indépendant de $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$ et donc également de $Y_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$ et comme $\frac{(n-2)S^2}{\sigma^2} \rightsquigarrow \chi^2(n-2)$:

$$\frac{Y_0 - (\widehat{\beta}_0 + \widehat{\beta}_1 x_0)}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}} \rightsquigarrow t(n-2).$$

Cette v.a. peut être encadrée par $[-t_{0,975}^{(n-2)}, t_{0,975}^{(n-2)}]$ avec probabilité 0,95 ce qui conduit immédiatement à l'intervalle de prédiction donné dans l'énoncé.
Notons que l'intervalle est d'autant plus précis que x_0 est proche de $\overline{x}$.

Exercice 11.5

Soit $\widetilde{\beta}_1 = \sum_{i=1}^n a_i Y_i$ un estimateur linéaire quelconque de β_1. Comme $E(\widetilde{\beta}_1) = \sum_{i=1}^n a_i E(Y_i) = \sum_{i=1}^n a_i(\beta_0 + \beta_1 x_i)$ il est sans biais, quels que soient les x_i, si $\sum_{i=1}^n a_i = 0$ et $\sum_{i=1}^n a_i x_i = 1$. Sa variance est $\sum_{i=1}^n a_i^2 V(Y_i) = \sigma^2 \sum_{i=1}^n a_i^2$.

L'estimateur de variance minimale parmi les estimateurs sans biais est tel que $\sum_{i=1}^n a_i^2$ soit minimal sous les contraintes $\sum_{i=1}^n a_i = 0$ et $\sum_{i=1}^n a_i x_i = 1$.

Résolvons ce problème par le multiplicateur de Lagrange, soit en minimisant la fonction de $(c_1, c_2, \cdots, c_n, \lambda_1, \lambda_2)$: $\sum_{i=1}^n a_i^2 - \lambda_1 \sum_{i=1}^n a_i - \lambda_2(\sum_{i=1}^n a_i x_i - 1)$.

On résout le système suivant annulant les $n+2$ dérivées partielles :

$$\left\{\begin{array}{l} 2a_i - \lambda_1 - \lambda_2 x_i = 0 \quad (n \text{ équations pour } i = 1, \cdots, n) \\ \sum_{i=1}^n a_i = 0 \\ \sum_{i=1}^n a_i x_i = 1 \end{array}\right. .$$

La somme des n premières équations donne $-n\lambda_1 - \lambda_2 \sum_{i=1}^n x_i = 0$. En multipliant chacune d'elles par x_i et en sommant on obtient encore $2 - \lambda_1 \sum_{i=1}^n x_i - \lambda_2 \sum_{i=1}^n x_i^2 = 0$. Ces deux équations permettre de déterminer λ_1 et λ_2 :

$$\left\{\begin{array}{c} \lambda_1 = -\lambda_2 \overline{x} \\ 2 + \lambda_2(\overline{x}\sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2) = 0 \end{array}\right. \iff \left\{\begin{array}{c} \lambda_2 = \frac{2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ \lambda_1 = -\frac{2\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \end{array}\right. .$$

Puis, en substituant dans les n premières équations du système initial :

$$2a_i + \frac{2\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} - \frac{2x_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = 0$$

soit finalement :

$$a_i = \frac{x_i - \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} .$$

Ainsi l'estimateur de variance minimale est :

$$\widetilde{\beta}_1^* = \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

puisque $\sum_{i=1}^n (x_i - \overline{x})\overline{Y} = 0$. Cet estimateur est bien celui des moindres carrés.

Exercice 11.6

La log-vraisemblance a été vue en section 11.2.2 :

$$\ln L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}(\ln 2\pi + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Le dénominateur du RVG est obtenu avec les estimations du MV de β_0, β_1 et σ^2 établis dans cette section :

$$\begin{aligned} \ln L(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\sigma}^2) &= -\frac{n}{2}(\ln 2\pi + \ln \widehat{\sigma}^2) - \frac{1}{2\widehat{\sigma}^2} \sum_{i=1}^n [y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2 \\ &= -\frac{n}{2}(\ln 2\pi + \ln \widehat{\sigma}^2) - \frac{n}{2} \end{aligned}$$

car $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2$.

Pour le numérateur, on se place sous $H_0 : \beta_1 = 0$, c'est-à-dire que les Y_i sont i.i.d. d'espérance β_0 et de variance σ^2 dont les estimations du MV sont respectivement (voir exemple 6.20) $\overline{y}$ et $\widetilde{s}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \overline{y})^2$. La log-vraisemblance maximale est :

$$\ln L(\widehat{\beta}_0, 0, \widetilde{s}^2) = -\frac{n}{2}(\ln 2\pi + \ln \widetilde{s}^2) - \frac{n}{2}.$$

Finalement :

$$\ln RVG = -\frac{n}{2} \ln \frac{\widetilde{s}^2}{\widehat{\sigma}^2}. = -\frac{n}{2} \ln \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n\widehat{\sigma}^2}.$$

Or la réalisation de la statistique F est (voir section 11.2.4) :

$$f = \frac{\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{s^2} = \frac{(n-2)\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2}{n\widehat{\sigma}^2},$$

S^2 ayant été défini comme égal à $n\widehat{\sigma}^2/(n-2)$. Par la décomposition de la somme des carrés totale on a :

$$\sum_{i=1}^n (\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^n (y_i - \overline{y})^2 - \sum_{i=1}^n (y_i - \widehat{y}_i)^2 = \sum_{i=1}^n (y_i - \overline{y})^2 - n\widehat{\sigma}^2$$

d'où :
$$f = (n-2)\left[\frac{\sum_{i=1}^n (y_i - \overline{y})^2}{n\widehat{\sigma}^2} - 1\right]$$

et :
$$\ln RVG = -\frac{n}{2} \ln(\frac{f}{n-2} + 1)\,.$$

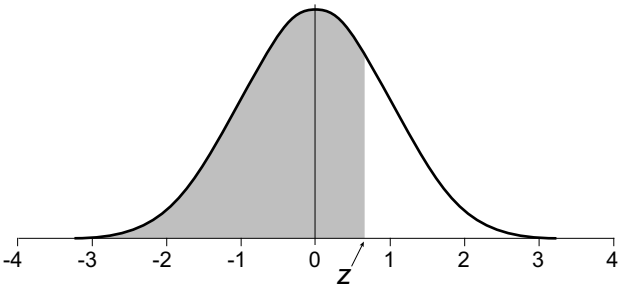
Le RVG est une fonction décroissante de f et la région de rejet de la forme $-2\ln RVG < k$ correspond à une région $f < k'$ du test classique. Toutefois la loi asymptotique de $-2\ln RVG$ n'est pas en correspondance avec la loi $F(1, n-2)$ du test classique.

Tables

Loi de Gauss
Loi de student
Loi de Fisher
Loi du khi-deux

Ces tables ont été reproduites avec l'aimable autorisation de Denis Labelle et Alain Latour, de leur ouvrage polycopié *Statistique Inférentielle*, Université du Québec à Montréal (UQAM)

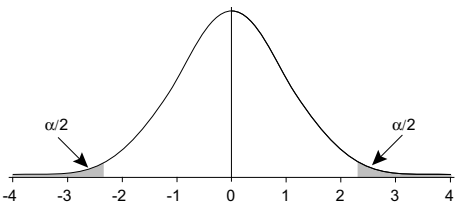
Loi normale



Loi $\mathcal{N}(0;1)$ *: Valeur de* $\Pr\{\mathcal{N}(0;1) \leq z\}$ *en fonction de* z

z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
0,0	,5000	,5040	,5080	,5120	,5160	,5199	,5239	,5279	,5319	,5359
0,1	,5398	,5438	,5478	,5517	,5557	,5596	,5636	,5675	,5714	,5753
0,2	,5793	,5832	,5871	,5910	,5948	,5987	,6026	,6064	,6103	,6141
0,3	,6179	,6217	,6255	,6293	,6331	,6368	,6406	,6443	,6480	,6517
0,4	,6554	,6591	,6628	,6664	,6700	,6736	,6772	,6808	,6844	,6879
0,5	,6915	,6950	,6985	,7019	,7054	,7088	,7123	,7157	,7190	,7224
0,6	,7257	,7291	,7324	,7357	,7389	,7422	,7454	,7486	,7517	,7549
0,7	,7580	,7611	,7642	,7673	,7704	,7734	,7764	,7794	,7823	,7852
0,8	,7881	,7910	,7939	,7967	,7995	,8023	,8051	,8078	,8106	,8133
0,9	,8159	,8186	,8212	,8238	,8264	,8289	,8315	,8340	,8365	,8389
1,0	,8413	,8438	,8461	,8485	,8508	,8531	,8554	,8577	,8599	,8621
1,1	,8643	,8665	,8686	,8708	,8729	,8749	,8770	,8790	,8810	,8830
1,2	,8849	,8869	,8888	,8907	,8925	,8944	,8962	,8980	,8997	,9015
1,3	,9032	,9049	,9066	,9082	,9099	,9115	,9131	,9147	,9162	,9177
1,4	,9192	,9207	,9222	,9236	,9251	,9265	,9279	,9292	,9306	,9319
1,5	,9332	,9345	,9357	,9370	,9382	,9394	,9406	,9418	,9429	,9441
1,6	,9452	,9463	,9474	,9484	,9495	,9505	,9515	,9525	,9535	,9545
1,7	,9554	,9564	,9573	,9582	,9591	,9599	,9608	,9616	,9625	,9633
1,8	,9641	,9649	,9656	,9664	,9671	,9678	,9686	,9693	,9699	,9706
1,9	,9713	,9719	,9726	,9732	,9738	,9744	,9750	,9756	,9761	,9767
2,0	,9772	,9778	,9783	,9788	,9793	,9798	,9803	,9808	,9812	,9817
2,1	,9821	,9826	,9830	,9834	,9838	,9842	,9846	,9850	,9854	,9857
2,2	,9861	,9864	,9868	,9871	,9875	,9878	,9881	,9884	,9887	,9890
2,3	,9893	,9896	,9898	,9901	,9904	,9906	,9909	,9911	,9913	,9916
2,4	,9918	,9920	,9922	,9925	,9927	,9929	,9931	,9932	,9934	,9936
2,5	,9938	,9940	,9941	,9943	,9945	,9946	,9948	,9949	,9951	,9952
2,6	,9953	,9955	,9956	,9957	,9959	,9960	,9961	,9962	,9963	,9964
2,7	,9965	,9966	,9967	,9968	,9969	,9970	,9971	,9972	,9973	,9974
2,8	,9974	,9975	,9976	,9977	,9977	,9978	,9979	,9979	,9980	,9981
2,9	,9981	,9982	,9982	,9983	,9984	,9984	,9985	,9985	,9986	,9986
3,0	,9987	,9987	,9987	,9988	,9988	,9989	,9989	,9989	,9990	,9990
3,1	,9990	,9991	,9991	,9991	,9992	,9992	,9992	,9992	,9993	,9993
3,2	,9993	,9993	,9994	,9994	,9994	,9994	,9994	,9995	,9995	,9995

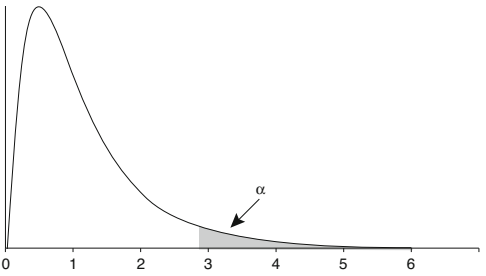
Loi de Student



α

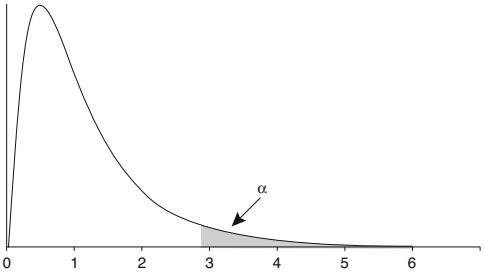
ν	0,500	0,200	0,100	0,050	0,020	0,010
1	1,000	3,078	6,314	12,706	31,821	63,656
2	0,816	1,886	2,920	4,303	6,965	9,925
3	0,765	1,638	2,353	3,182	4,541	5,841
4	0,741	1,533	2,132	2,776	3,747	4,604
5	0,727	1,476	2,015	2,571	3,365	4,032
6	0,718	1,440	1,943	2,447	3,143	3,707
7	0,711	1,415	1,895	2,365	2,998	3,499
8	0,706	1,397	1,860	2,306	2,896	3,355
9	0,703	1,383	1,833	2,262	2,821	3,250
10	0,700	1,372	1,812	2,228	2,764	3,169
11	0,697	1,363	1,796	2,201	2,718	3,106
12	0,695	1,356	1,782	2,179	2,681	3,055
13	0,694	1,350	1,771	2,160	2,650	3,012
14	0,692	1,345	1,761	2,145	2,624	2,977
15	0,691	1,341	1,753	2,131	2,602	2,947
16	0,690	1,337	1,746	2,120	2,583	2,921
17	0,689	1,333	1,740	2,110	2,567	2,898
18	0,688	1,330	1,734	2,101	2,552	2,878
19	0,688	1,328	1,729	2,093	2,539	2,861
20	0,687	1,325	1,725	2,086	2,528	2,845
21	0,686	1,323	1,721	2,080	2,518	2,831
22	0,686	1,321	1,717	2,074	2,508	2,819
23	0,685	1,319	1,714	2,069	2,500	2,807
24	0,685	1,318	1,711	2,064	2,492	2,797
25	0,684	1,316	1,708	2,060	2,485	2,787
26	0,684	1,315	1,706	2,056	2,479	2,779
27	0,684	1,314	1,703	2,052	2,473	2,771
28	0,683	1,313	1,701	2,048	2,467	2,763
29	0,683	1,311	1,699	2,045	2,462	2,756
30	0,683	1,310	1,697	2,042	2,457	2,750
50	0,679	1,299	1,676	2,009	2,403	2,678
70	0,678	1,294	1,667	1,994	2,381	2,648
90	0,677	1,291	1,662	1,987	2,368	2,632
∞	0,674	1,282	1,645	1,960	2,326	2,576

Loi de Fisher



$\Pr(F_{\nu_1,\nu_2} > c) = \mathbf{0,05}$

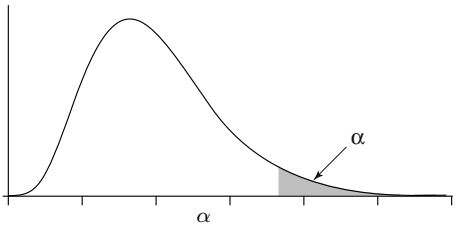
	ν_2											
ν_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**
1	161	18,5	10,1	7,71	6,61	5,99	5,59	5,32	5,12	4,96	4,75	4,54
2	199	19,0	9,55	6,94	5,79	5,14	4,74	4,46	4,26	4,10	3,89	3,68
3	216	19,2	9,28	6,59	5,41	4,76	4,35	4,07	3,86	3,71	3,49	3,29
4	225	19,2	9,12	6,39	5,19	4,53	4,12	3,84	3,63	3,48	3,26	3,06
5	230	19,3	9,01	6,26	5,05	4,39	3,97	3,69	3,48	3,33	3,11	2,90
6	234	19,3	8,94	6,16	4,95	4,28	3,87	3,58	3,37	3,22	3,00	2,79
7	237	19,4	8,89	6,09	4,88	4,21	3,79	3,50	3,29	3,14	2,91	2,71
8	239	19,4	8,85	6,04	4,82	4,15	3,73	3,44	3,23	3,07	2,85	2,64
9	241	19,4	8,81	6,00	4,77	4,10	3,68	3,39	3,18	3,02	2,80	2,59
10	242	19,4	8,79	5,96	4,74	4,06	3,64	3,35	3,14	2,98	2,75	2,54
11	243	19,4	8,76	5,94	4,70	4,03	3,60	3,31	3,10	2,94	2,72	2,51
12	244	19,4	8,74	5,91	4,68	4,00	3,57	3,28	3,07	2,91	2,69	2,48
13	245	19,4	8,73	5,89	4,66	3,98	3,55	3,26	3,05	2,89	2,66	2,45
14	245	19,4	8,71	5,87	4,64	3,96	3,53	3,24	3,03	2,86	2,64	2,42
15	246	19,4	8,70	5,86	4,62	3,94	3,51	3,22	3,01	2,85	2,62	2,40
16	246	19,4	8,69	5,84	4,60	3,92	3,49	3,20	2,99	2,83	2,60	2,38
17	247	19,4	8,68	5,83	4,59	3,91	3,48	3,19	2,97	2,81	2,58	2,37
18	247	19,4	8,67	5,82	4,58	3,90	3,47	3,17	2,96	2,80	2,57	2,35
19	248	19,4	8,67	5,81	4,57	3,88	3,46	3,16	2,95	2,79	2,56	2,34
20	248	19,4	8,66	5,80	4,56	3,87	3,44	3,15	2,94	2,77	2,54	2,33
21	248	19,4	8,65	5,79	4,55	3,86	3,43	3,14	2,93	2,76	2,53	2,32
22	249	19,5	8,65	5,79	4,54	3,86	3,43	3,13	2,92	2,75	2,52	2,31
23	249	19,5	8,64	5,78	4,53	3,85	3,42	3,12	2,91	2,75	2,51	2,30
24	249	19,5	8,64	5,77	4,53	3,84	3,41	3,12	2,90	2,74	2,51	2,29
25	249	19,5	8,63	5,77	4,52	3,83	3,40	3,11	2,89	2,73	2,50	2,28
26	249	19,5	8,63	5,76	4,52	3,83	3,40	3,10	2,89	2,72	2,49	2,27
27	250	19,5	8,63	5,76	4,51	3,82	3,39	3,10	2,88	2,72	2,48	2,27
28	250	19,5	8,62	5,75	4,50	3,82	3,39	3,09	2,87	2,71	2,48	2,26
29	250	19,5	8,62	5,75	4,50	3,81	3,38	3,08	2,87	2,70	2,47	2,25
30	250	19,5	8,62	5,75	4,50	3,81	3,38	3,08	2,86	2,70	2,47	2,25
32	250	19,5	8,61	5,74	4,49	3,80	3,37	3,07	2,85	2,69	2,46	2,24
34	251	19,5	8,61	5,73	4,48	3,79	3,36	3,06	2,85	2,68	2,45	2,23
36	251	19,5	8,60	5,73	4,47	3,79	3,35	3,06	2,84	2,67	2,44	2,22
38	251	19,5	8,60	5,72	4,47	3,78	3,35	3,05	2,83	2,67	2,43	2,21
40	251	19,5	8,59	5,72	4,46	3,77	3,34	3,04	2,83	2,66	2,43	2,20
45	251	19,5	8,59	5,71	4,45	3,76	3,33	3,03	2,81	2,65	2,41	2,19
50	252	19,5	8,58	5,70	4,44	3,75	3,32	3,02	2,80	2,64	2,40	2,18
55	252	19,5	8,58	5,69	4,44	3,75	3,31	3,01	2,79	2,63	2,39	2,17
60	252	19,5	8,57	5,69	4,43	3,74	3,30	3,01	2,79	2,62	2,38	2,16
65	252	19,5	8,57	5,68	4,43	3,73	3,30	3,00	2,78	2,61	2,38	2,15
70	252	19,5	8,57	5,68	4,42	3,73	3,29	2,99	2,78	2,61	2,37	2,15
80	253	19,5	8,56	5,67	4,41	3,72	3,29	2,99	2,77	2,60	2,36	2,14



$\Pr\left(F_{\nu_1,\nu_2} > c\right) = \mathbf{0,05}$

	ν_2											
ν_1	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**40**	**50**	**60**	**100**	**200**
1	4,41	4,35	4,30	4,26	4,23	4,20	4,17	4,08	4,03	4,00	3,94	3,89
2	3,55	3,49	3,44	3,40	3,37	3,34	3,32	3,23	3,18	3,15	3,09	3,04
3	3,16	3,10	3,05	3,01	2,98	2,95	2,92	2,84	2,79	2,76	2,70	2,65
4	2,93	2,87	2,82	2,78	2,74	2,71	2,69	2,61	2,56	2,53	2,46	2,42
5	2,77	2,71	2,66	2,62	2,59	2,56	2,53	2,45	2,40	2,37	2,31	2,26
6	2,66	2,60	2,55	2,51	2,47	2,45	2,42	2,34	2,29	2,25	2,19	2,14
7	2,58	2,51	2,46	2,42	2,39	2,36	2,33	2,25	2,20	2,17	2,10	2,06
8	2,51	2,45	2,40	2,36	2,32	2,29	2,27	2,18	2,13	2,10	2,03	1,98
9	2,46	2,39	2,34	2,30	2,27	2,24	2,21	2,12	2,07	2,04	1,97	1,93
10	2,41	2,35	2,30	2,25	2,22	2,19	2,16	2,08	2,03	1,99	1,93	1,88
11	2,37	2,31	2,26	2,22	2,18	2,15	2,13	2,04	1,99	1,95	1,89	1,84
12	2,34	2,28	2,23	2,18	2,15	2,12	2,09	2,00	1,95	1,92	1,85	1,80
13	2,31	2,25	2,20	2,15	2,12	2,09	2,06	1,97	1,92	1,89	1,82	1,77
14	2,29	2,22	2,17	2,13	2,09	2,06	2,04	1,95	1,89	1,86	1,79	1,74
15	2,27	2,20	2,15	2,11	2,07	2,04	2,01	1,92	1,87	1,84	1,77	1,72
16	2,25	2,18	2,13	2,09	2,05	2,02	1,99	1,90	1,85	1,82	1,75	1,69
17	2,23	2,17	2,11	2,07	2,03	2,00	1,98	1,89	1,83	1,80	1,73	1,67
18	2,22	2,15	2,10	2,05	2,02	1,99	1,96	1,87	1,81	1,78	1,71	1,66
19	2,20	2,14	2,08	2,04	2,00	1,97	1,95	1,85	1,80	1,76	1,69	1,64
20	2,19	2,12	2,07	2,03	1,99	1,96	1,93	1,84	1,78	1,75	1,68	1,62
21	2,18	2,11	2,06	2,01	1,98	1,95	1,92	1,83	1,77	1,73	1,66	1,61
22	2,17	2,10	2,05	2,00	1,97	1,93	1,91	1,81	1,76	1,72	1,65	1,60
23	2,16	2,09	2,04	1,99	1,96	1,92	1,90	1,80	1,75	1,71	1,64	1,58
24	2,15	2,08	2,03	1,98	1,95	1,91	1,89	1,79	1,74	1,70	1,63	1,57
25	2,14	2,07	2,02	1,97	1,94	1,91	1,88	1,78	1,73	1,69	1,62	1,56
26	2,13	2,07	2,01	1,97	1,93	1,90	1,87	1,77	1,72	1,68	1,61	1,55
27	2,13	2,06	2,00	1,96	1,92	1,89	1,86	1,77	1,71	1,67	1,60	1,54
28	2,12	2,05	2,00	1,95	1,91	1,88	1,85	1,76	1,70	1,66	1,59	1,53
29	2,11	2,05	1,99	1,95	1,91	1,88	1,85	1,75	1,69	1,66	1,58	1,52
30	2,11	2,04	1,98	1,94	1,90	1,87	1,84	1,74	1,69	1,65	1,57	1,52
32	2,10	2,03	1,97	1,93	1,89	1,86	1,83	1,73	1,67	1,64	1,56	1,50
34	2,09	2,02	1,96	1,92	1,88	1,85	1,82	1,72	1,66	1,62	1,55	1,49
36	2,08	2,01	1,95	1,91	1,87	1,84	1,81	1,71	1,65	1,61	1,54	1,48
38	2,07	2,00	1,95	1,90	1,86	1,83	1,80	1,70	1,64	1,60	1,52	1,47
40	2,06	1,99	1,94	1,89	1,85	1,82	1,79	1,69	1,63	1,59	1,52	1,46
45	2,05	1,98	1,92	1,88	1,84	1,80	1,77	1,67	1,61	1,57	1,49	1,43
50	2,04	1,97	1,91	1,86	1,82	1,79	1,76	1,66	1,60	1,56	1,48	1,41
55	2,03	1,96	1,90	1,85	1,81	1,78	1,75	1,65	1,59	1,55	1,46	1,40
60	2,02	1,95	1,89	1,84	1,80	1,77	1,74	1,64	1,58	1,53	1,45	1,39
65	2,01	1,94	1,88	1,83	1,79	1,76	1,73	1,63	1,57	1,52	1,44	1,37
70	2,00	1,93	1,88	1,83	1,79	1,75	1,72	1,62	1,56	1,52	1,43	1,36
80	1,99	1,92	1,86	1,82	1,78	1,74	1,71	1,61	1,54	1,50	1,41	1,35

Loi du khi-deux



ν	0,995	0,975	0,95	0,9	0,1	0,05	0,025	0,005
1	0,000	0,001	0,004	0,016	2,706	3,841	5,024	7,879
2	0,010	0,051	0,103	0,211	4,605	5,991	7,378	10,597
3	0,072	0,216	0,352	0,584	6,251	7,815	9,348	12,838
4	0,207	0,484	0,711	1,064	7,779	9,488	11,143	14,860
5	0,412	0,831	1,145	1,610	9,236	11,070	12,832	16,750
6	0,676	1,237	1,635	2,204	10,645	12,592	14,449	18,548
7	0,989	1,690	2,167	2,833	12,017	14,067	16,013	20,278
8	1,344	2,180	2,733	3,490	13,362	15,507	17,535	21,955
9	1,735	2,700	3,325	4,168	14,684	16,919	19,023	23,589
10	2,156	3,247	3,940	4,865	15,987	18,307	20,483	25,188
11	2,603	3,816	4,575	5,578	17,275	19,675	21,920	26,757
12	3,074	4,404	5,226	6,304	18,549	21,026	23,337	28,300
13	3,565	5,009	5,892	7,041	19,812	22,362	24,736	29,819
14	4,075	5,629	6,571	7,790	21,064	23,685	26,119	31,319
15	4,601	6,262	7,261	8,547	22,307	24,996	27,488	32,801
16	5,142	6,908	7,962	9,312	23,542	26,296	28,845	34,267
17	5,697	7,564	8,672	10,085	24,769	27,587	30,191	35,718
18	6,265	8,231	9,390	10,865	25,989	28,869	31,526	37,156
19	6,844	8,907	10,117	11,651	27,204	30,144	32,852	38,582
20	7,434	9,591	10,851	12,443	28,412	31,410	34,170	39,997
21	8,034	10,283	11,591	13,240	29,615	32,671	35,479	41,401
22	8,643	10,982	12,338	14,041	30,813	33,924	36,781	42,796
23	9,260	11,689	13,091	14,848	32,007	35,172	38,076	44,181
24	9,886	12,401	13,848	15,659	33,196	36,415	39,364	45,558
25	10,520	13,120	14,611	16,473	34,382	37,652	40,646	46,928
26	11,160	13,844	15,379	17,292	35,563	38,885	41,923	48,290
27	11,808	14,573	16,151	18,114	36,741	40,113	43,195	49,645
28	12,461	15,308	16,928	18,939	37,916	41,337	44,461	50,994
29	13,121	16,047	17,708	19,768	39,087	42,557	45,722	52,335
30	13,787	16,791	18,493	20,599	40,256	43,773	46,979	53,672
40	20,707	24,433	26,509	29,051	51,805	55,758	59,342	66,766
50	27,991	32,357	34,764	37,689	63,167	67,505	71,420	79,490
60	35,534	40,482	43,188	46,459	74,397	79,082	83,298	91,952
70	43,275	48,758	51,739	55,329	85,527	90,531	95,023	104,215
80	51,172	57,153	60,391	64,278	96,578	101,879	106,629	116,321
90	59,196	65,647	69,126	73,291	107,565	113,145	118,136	128,299
100	67,328	74,222	77,929	82,358	118,498	124,342	129,561	140,170

Bibliographie

Agresti A (2002) *Categorical data Analysis*. Second edition, Wiley, New York

Bosq D, Lecoutre JP (1987) *Théorie de l'estimation fonctionnelle.* Economica, Paris

Chap TL (1998) *Applied Categorical Analysis.* Wiley, New York

Chernoff H, Lehmann EL (1954) The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics* **25 :** 579-86

Cleveland WS (1979) Robust Locally Weighted Regression. *Journal of the American Statistical Association* **74 :** 829-36

Collomb G (1977) Quelques propriétés de la méthode du noyau pour l'estimation non-paramétrique de la régression en un point fixé. *Comptes Rendus de l'Académie des Sciences de Paris* tome 285, série A : 289-92

Cox DR, Hinkley DV (1979) *Theoretical Statistics.* Chapmann and Hall, London

Cramer H (1946) *Mathematical Methods of Statistics.* Princeton University Press

Davison AC, Hinkley DV (1997) *Bootstrap Methods and their Application.* Cambridge University Press

Deheuvels P (1977) Estimation Non Paramétrique de la Densité par histogrammes généralisés. *Revue de Statistique Appliquée* vol. XV : 5-42

Devroye L, Györfi L (1985) *Nonparametric Density Estimation : The L1 View.* Wiley, New York

Droesbeke JJ, Fine J éd. (1996) *Inférence non paramétrique, les statistiques de rangs.* Editions de l'Université de Bruxelles et Ellipses

Droesbeke JJ, Lejeune M, Saporta G éd. (2004) *Données Catégorielles.* Technip, Paris

Dodge Y (1999) *Analyse de régression appliquée.* Dunod, Paris

Efron B (1979) Bootstrap methods : another look at the jackknife. *Annals of Statistics* **7** : 1-26

Fan J (1993) Local Linear Regression Smoothers and their Minimax Efficiency. *Annals of Statistics* **21**, 1 : 196-216

Friedman D, Diaconis P (1981) On the Histogram as a Density Estimator : L2 Theory. *Zeitung fuer Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57** : 453-76

Gasser T, Müller HG (1984) Estimating Regression Functions and their Derivatives by the Kernel Method. *Scandinavian Journal of Statistics* **11** : 171-85

Gibbons JD (1985) *Nonparametric Statistical Inference,* Second edition. Marcel Dekker, New York

Haerdle W (1990) *Applied Nonparametric Regression.* Cambridge University Press

Hodges JL, Lehmann EL (1956) The efficiency of some non parametric competitors of the *t*-test. *Annals of Mathematical Statistics* **27 :** 324-55

Hodges JL, Lehmann EL (1963) Estimates of location based on rank tests. *Annals of Mathematical Statistics* **34 :** 598-611

Hosmer DW, Lemeshow S (2000) *Applied Logistic Regression,* Second edition. Wiley, New York

Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *Annals of Mathematical Statistics* **27** : 887-906

Lecoutre JP, Tassi P (1987) *Statistique non paramétrique et robustesse.* Economica, Paris

Lehmann EL (1975) *Nonparametrics : Statistical Methods based on Ranks.* Holden-Day, San Francisco

Lehmann EL (1986) *Testing Statistical Hypotheses,* Second edition. Wiley, New York

Lehmann EL, Casella G. (1998) *Theory of Point Estimation.* Springer-Verlag, New York

Lejeune M (1982) Estimation de densité à noyau variable. *Rapport technique No 1, Projet No 2.843-0.80 du Fonds National Suisse de la Recherche Scientifique «Développement et implémentation de méthodes d'estimation non paramétrique»*

Lejeune M (1983) Estimation non-paramétrique multivariée par noyaux. *Rapport technique No 3, Projet No 2.843-0.80 du FNRS*

Lejeune M (1984) Optimization in non Parametric Regression. *Compstat 84, Proceedings in Computational Statistics.* Physica-Verlag, Wien *:* 421-26

Lejeune M (1985) Estimation Non Paramétrique par Noyaux : Régression Polynomiale Mobile. *Revue de Statistique Appliquée* vol. XXXIII, N^o 3 : 43-67

Lejeune M, Sarda P (1992) Smooth estimators of distribution and density functions. *Computational Statistics & Data Analysis* **14** : 451-71

Mann HB, Whitney DR (1947) On a test of whether one of the two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* **18** : 50-60

Marron JS (1987) A Comparison of Cross-Validation Techniques in Density Estimation. *Annals of Statistics* **15** : 152-62

Mosteller F, Tukey JW (1977) *Data analysis and regression, a second course in statistics.* Addison-Wesley

Nadaraya EA (1964) On Estimating Regression. *Theory of Probability and its Applications* **9** : 141-42

Parzen E (1962) On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics* **33** : 1065-76

Quenouille M (1956) Notes on biais in estimation. *Biometrika* **43** : 353-60

Rosenblatt M (1956) Remarks on Some Nonparamatric Estimates of a Density Function. *Annals of Mathematical Statistics* **27** : 832-5

Saporta G (1990) *Probabilités, Analyse des Données et Statistique.* Technip, Paris

Seber GAF (1977) *Linear Regression Analysis.* Wiley, New York

Shao J, Tu D (1995) *The Jackknife and Bootstrap.* Springer-Verlag, New York

Shao J (1999) *Mathematical Statistics.* Springer-Verlag, New York

Shapiro S, Wilk M (1965) An analysis of variance test for normality. *Biometrika* **52** : 591-611

Silverman BW (1987) The bootstrap : To smooth or not to smooth ? *Biometrika* **74**,3 : 469-79

Simonoff JS (1996) *Smoothing Techniques in Statistics.* Springer-Verlag, New York

Watson GS (1964) Smooth Regression Analysis. *Sankhya A* **26** : 359-72

Wilcoxon F (1945) Individual comparisons of grouped data by ranking methods. *Biometrics Bulletin* **1** : 80-3

Index

E

F

G

H

I

J

K

L

N

O

P

Q

R

S

T

U

V