MDI221
Statistiques élémentaires

Pascal Bianchi, Philippe Ciblat

27 septembre 2023

Table des matières

1. Bases mathématiques

1.1. Algèbre linéaire

1.1.1. Espace vectoriel

On appelle espace vectoriel un ensemble $\mathcal{E}$ vérifiant les propriétés suivantes :

- il existe une opération, notée $+$, qui est interne (deux éléments additionnés de $\mathcal{E}$ restent dans $\mathcal{E}$), associative, possède un élément neutre, noté 0, et est symétrique (si $A + B = 0$ alors $B + A = 0$ aussi)
- il existe une opération, notée ., qui est externe. Soit un élément A de $\mathcal{E}$ et un élément λ d'un autre ensemble $\mathcal{K}$ (possédant certaines propriétés que nous ne mentionnerons pas ici), alors $\lambda.A$ est aussi dans $\mathcal{E}$. Et il y a une distributivité entre les opérations interne et externe, c-à-d, $\lambda(A + B) = \lambda A + \lambda B$ et $(\lambda + \mu)A = \lambda A + \mu A$.

L'exemple le plus courant est $\mathbb{R}^m$ pour lequel quand on additionne deux éléments de $\mathbb{R}^m$ on reste dans $\mathbb{R}^m$ et quand on multiple un vecteur par un réel (ici, on a $\mathcal{K} = \mathbb{R}$), on reste dans $\mathbb{R}^m$ aussi. De plus les lois de multiplication et d'addition ont bien les propriétés voulues, comme celle de la distributivité.

De plus on peut définir un sous-espace vectoriel à $\mathcal{E}$, noté $\mathcal{F}$, dont tous les éléments sont dans $\mathcal{E}$ mais quand on additionne deux éléments de $\mathcal{F}$, on reste dans $\mathcal{F}$ et quand on multiplie un élément de $\mathcal{F}$ par un point de $\mathcal{K}$ on reste dans $\mathcal{F}$.

Comme exemple, on peut penser à $\mathcal{F} = \mathbb{R}.\mathbf{1}_m$ avec $\mathcal{E} = \mathbb{R}^m$ où $\mathbf{1}_m$ est le vecteur de taille m composé uniquement de 1.

Dans la suite de ce polycopié, nous ne travaillerons qu'avec $\mathcal{E} = \mathbb{R}^m$.

1.1.2. Matrices

On considère un système de m équations avec d inconnues réelles à résoudre.

$$\left\{ \begin{array}{lcl} a_{1,1}x_1 + \cdots + a_{1,d}x_d & = & b_1 \\ \ldots & = & \ldots \\ a_{m,1}x_1 + \cdots + a_{m,d}x_d & = & b_m \end{array} \right.$$

On appelle matrice A, le tableau suivant

$$A = \begin{pmatrix} a_{1,1} \cdots\cdots a_{1,d} \\ \vdots \\ a_{m,1} \cdots\cdots a_{m,d} \end{pmatrix}$$

et le système d'équations précédent s'écrit alors

$$A \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

Image d'une matrice

Une collection de vecteurs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ de $\mathbb{R}^m$ forme une *famille libre* si la propriété suivante est vraie :

$$\forall x_1, \ldots, x_d,\ x_1\boldsymbol{a}_1 + \cdots + x_d\boldsymbol{a}_d = 0 \text{ implique que } x_1 = \cdots = x_d = 0\,.$$

Cela signifie que dans une famille libre, aucun vecteur ne peut s'écrire comme combinaison linéaire des autres. Si $\mathcal{E}$ est un sous-espace vectoriel de $\mathbb{R}^m$, on appelle *base de* $\mathcal{E}$ toute famille libre de $\mathcal{E}$ telle que, en outre, n'importe quel élément de $\mathcal{E}$ puisse s'écrire comme combinaison linéaire des vecteurs de la base. La *dimension* de l'espace vectoriel $\mathcal{E}$ est le nombre de vecteurs nécessaires pour constituer une base.

Soit A une matrice $m \times d$. L'image de A est l'ensemble :

$$\mathrm{Im}(A) = \{y \in \mathbb{R}^m \,:\, \exists x \in \mathbb{R}^d, y = Ax\}\,.$$

C'est un sous-espace vectoriel de $\mathbb{R}^m$. Plus exactement, si on note $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ les colonnes de la matrice A, soit :

$$A = \left(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d\right)\,, \tag{1.1}$$

alors $\mathrm{Im}(A)$ est l'ensemble des combinaison linéraires de ces vecteurs. Comme il est important d'avoir compris cela, nous démontrons ce point :

$$\begin{aligned} y \in \mathrm{Im}(A) \quad &\Leftrightarrow \quad \exists x,\ y = Ax \\ &\Leftrightarrow \quad \exists x,\ y = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d] \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \\ &\Leftrightarrow \quad \exists x,\ y = x_1\boldsymbol{a}_1 + \cdots + x_d\boldsymbol{a}_d \\ &\Leftrightarrow \quad y \text{ est une combinaison linéaire des colonnes de } A\,. \end{aligned}$$

L'image $\mathrm{Im}(A)$ d'une matrice A est donc le *sous-espace-vectoriel engendré par les colonnes de A*. Le *rang* de A, noté $\mathrm{rang}(A)$ est la dimension de $\mathrm{Im}(A)$. Dans de nombreux cas pratiques, les d colonnes de A forment une famille libre, et par conséquent, forment une base de $\mathrm{Im}(A)$. Dans ce cas, $\mathrm{rang}(A) = d$. Mais dans d'autres cas, il peut y avoir de la redondance dans les colonnes de A (par exemple, une colonne est répétée, ou une colonne est la somme de deux autres). Dans ce cas de figure, il faudra moins de d vecteurs pour constituer une base de $\mathrm{Im}(A)$, et ainsi le rang de A sera plus petit que d.

Le noyau d'une matrice A est l'ensemble :

$$\ker(A) = \{x \in \mathbb{R}^d \,:\, Ax = 0\}\,.$$

C'est un sous-espace vectoriel de $\mathbb{R}^d$. Le théorème du rang stipule que :

$$\text{rang}(A) + \dim(\ker(A)) = d\,.$$

En particulier, on a l'équivalence suivante :

$$\text{rang}(A) = d \;\Leftrightarrow\; \ker(A) = \{0\} \;\Leftrightarrow\; \boldsymbol{a}_1, \ldots, \boldsymbol{a}_d \text{ est une famille libre.}$$

On rappelle que la transposée d'une matrice A est notée A^T. La transposition consiste à transformer les colonnes de A en lignes de A^T. On rappelle la règle : $(AB)^T = B^T A^T$. La transposée d'un vecteur colonne est un vecteur ligne.

On appelle trace(A) la somme des éléments diagonaux de A.

Matrices carrées

On dit qu'une matrice carrée $d \times d$ est

- inversible s'il existe une matrice, notée A^{-1} (et forcément unique), telle que $AA^{-1} = A^{-1}A = I_d$, où I_d est la matrice identité de taille $d \times d$. Cela revient à dire que $\ker(A) = \{0\}$, ou de manière équivalente que $\text{rang}(A) = d$, où de manière encore équivalente, que $\det(A) \neq 0$, où $\det(A)$ est le *déterminant* de A.
- orthogonale si en outre $A^T = A^{-1}$.
- admet une valeurs propre et un vecteur propre si il existe un réel λ (valeur propre) et un vecteur v de $\mathbb{R}^d$ tels que
$$Av = \lambda v.$$
En général, une matrice admet un ensemble de valeurs propres et de vecteurs propres.
- symétrique si $A = A^T$.

Quand une matrice est symétrique, nous avons quelques définitions supplémentaires et propriétés intéressantes.

- A est symétrique semi-définie positive, si en outre $x^T Ax \geq 0$ pour tout vecteur colonne x de taille d.
- A symétrique définie positive, si en outre $x^T Ax > 0$ dès que $x \neq 0$.
- Soient deux matrices A et B symétriques. On dira que A est supérieure à B, et on notera $A \succeq B$ si la matrice $(A - B)$ qui est symétrique est aussi semi-définie positive, c-à-d, si $x^T(A - B)x \geq 0 \Leftrightarrow x^T Ax \geq x^T Bx$ pour tout vecteur colonne x de taille d.

Un des résultats les plus importants d'algèbre linéaire est le suivant.

Toute matrice symétrique est diagonalisable dans une base orthogonale

c'est à dire que si A est symétrique, alors il existe une matrice orthogonale P et une matrice diagonale Λ, telles que

$$A = P\Lambda P^T . \tag{1.2}$$

Dans ce cas, les coefficients de la diagonale de Λ coïncident avec les valeurs propres de A. Les colonnes de P sont constituées des vecteurs propres. De plus

— Si A est symétrique inversible, alors toutes ces valeurs propres sont non-nulles et la matrice inverse vaut

$$A^{-1} = P\Lambda^{-1}P^T.$$

— Si A est symétrique semi-définie positive, alors les valeurs propres (les coefficients sur la diagonale de Λ) sont positives ou nulles.

— Si A est définie positive, ces valeurs propres sont strictement positives, et dans ce cas la matrice A est inversible.

Produit scalaire

Soient deux vecteurs x, y dans $\mathbb{R}^m$. On supposera toujours que les vecteurs sont des vecteurs-colonne, et on notera x_i, y_i leurs composantes respectives. Le produit scalaire de x et y est défini par :

$$\begin{aligned}\langle x, y\rangle &= \sum_{i=1}^{m} x_i y_i \\ &= x^T y\,.\end{aligned}$$

La norme (euclidienne) de x est définie par :

$$\|x\| = \sqrt{x^T x}\,.$$

1.2. Optimisation

1.2.1. Dérivées et minimiseurs

Si $f : \mathbb{R}^d \to \mathbb{R}$ est une fonction, on dit que x est un *minimiseur* de f si $f(y) \geq f(x)$ pour tout y. On note $\arg\min f$ l'ensemble des minimiseurs de f (c'est un ensemble, mais lorsque f admet un unique minimiseur, $\arg\min f$ est simplement un point de $\mathbb{R}^d$). Le gradient d'une fonction f en un point $x \in \mathbb{R}^d$ est le vecteur :

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix} .$$

Dans le cas unidimensionnel $d = 1$, le gradient est simplement la dérivée $f'(x)$ de la fonction. On peut aller un cran plus loin, et définir la *matrice hessienne* de f au point x. Il s'agit de la

matrice $d \times d$ définie par :

$$\mathrm{Hess}(f) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ \vdots & & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d^2} \end{pmatrix}$$

c'est à dire que le coefficient (i, j) de la matrice est $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. Cette matrice est symétrique, car on peut permuter l'ordre de dérivation entre x_i et x_j. Dans le cas unidimensionnel $d = 1$, le Hessien est simplement la dérivée-seconde $f''(x)$ de la fonction.

Un point x est un *point critique* de f si $\nabla f(x) = 0$. Tout minimiseur de f est un point critique. La réciproque n'est pas vraie en générale, mais elle est vraie si f est une fonction convexe. Une fonction f est dite *convexe* si la propriété suivante est satisfaite :

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

pour tout $t \in [0, 1]$ et tous points x et y dans $\mathbb{R}^d$. On notera que l'ensemble $\{tf(x)+(1-t)f(y), t \in [0, 1]\}$ est le segment de droite reliant $f(x)$ à $f(y)$.

Théorème 1.1. *Dans le cas où f est convexe, on a effectivement l'équivalence :*

$$x \in \arg\min f \;\Leftrightarrow\; \nabla f(x) = 0\,.$$

Cela signifie que pour trouver un minimiseur d'une fonction il suffit de chercher un point qui annule le gradient. De ce point de vue, les fonctions convexes sont sympathiques, car on a alors un outil simple pour rechercher ces minimiseurs. En pratique, il faudra savoir repérer qu'une fonction est convexe. Dans le cas unidimensionnel $d = 1$ (f est une fonction de $\mathbb{R} \to \mathbb{R}$), on se convainc facilement que f est convexe si et seulement sa dérivée est croissante (faire un dessin pour s'en convaincre). Autrement dit, f est convexe si et seulement si sa dérivée-seconde est partout positive ou nulle :

$$f : \mathbb{R} \to \mathbb{R} \text{ est convexe } \;\Leftrightarrow\; \forall x \in \mathbb{R},\; f''(x) \geq 0\,.$$

Ce résultat intuitif admet une généralisation dans $\mathbb{R}^d$, donnée par le théorème suivant :

Théorème 1.2. *Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est convexe si et seulement si* $\mathrm{Hess}(f)$ *est semi-définie positive.*

1.2.2. Projection orthogonale sur un sous-espace

Soit A la matrice $m \times d$ dont les colonnes sont données par $\boldsymbol{a}_1, \ldots \boldsymbol{a}_d$, comme dans l'équation (1.1). Noter que ces colonnes sont des vecteurs de $\mathbb{R}^m$.

Soit $z \in \mathbb{R}^m$ un point quelconque. On appelle *projeté de z sur* $\mathrm{Im}(A)$ le point, noté $\Pi_A(z)$, qui appartient à $\mathrm{Im}(A)$, et qui est le plus proche de z parmi tous les points de $\mathrm{Im}(A)$. En clair :

$$\Pi_A(z) = \arg \min_{y \in \mathrm{Im}(A)} \|y - z\|\,.$$

Théorème 1.3. *Supposons que* $\mathrm{rang}(A) = d$. *Alors :*

$$\Pi_A(z) = A(A^TA)^{-1}A^Tz\,.$$

La matrice $\Pi_A = A(A^TA)^{-1}A^T$ *est appelée le projecteur orthogonal sur* $\mathrm{Im}(A)$.

Démonstration. Avant toute chose, on remarque que la matrice A^TA est bien inversible. En effet, sous l'hypothèse que $\mathrm{rang}(A) = d$, le théorème du rang implique que $\ker(A) = \{0\}$. Donc, par l'exercice 1.1, $\ker(A^TA) = \{0\}$, ce qui implique que A^TA est bien inversible. Passons maintenant au reste de la preuve.

Appelons temporairement Π la matrice $A(A^TA)^{-1}A^T$. Soit $y \in \mathrm{Im}(A)$ un élément quelconque de l'image de A. Par définition, il existe un certain $x \in \mathbb{R}^d$ tel que $y = Ax$. Un petit calcul montre que :

$$\Pi y = \Pi Ax = A(A^TA)^{-1}A^TAx = Ax = y\,.$$

On peut donc écrire que pour tout $y \in \mathrm{Im}(A)$,

$$z - y = \Pi(z - y) + (I - \Pi)z\,.$$

où I est ici la matrice identité de taille $m \times m$. Ainsi, d'après l'exercice 1.5,

$$\|z - y\|^2 = \|\Pi(z - y)\|^2 + 2\langle \Pi(z - y), (I - \Pi)z\rangle + \|(I - \Pi)z\|^2\,.$$

En fait, le lecteur peut très facilement vérifier que

$$\Pi^2 = \Pi = \Pi^T\,.$$

Par conséquent, on montre facilement que le produit scalaire est nul :

$$\langle \Pi(z - y), (I - \Pi)z\rangle = (z - y)^T\Pi(I - \Pi)z = (z - y)^T(\Pi - \Pi^2)z = 0\,.$$

Donc, on peut simplifier :

$$\|z - y\|^2 = \|\Pi(z - y)\|^2 + \|(I - \Pi)z\|^2\,.$$

On a montré que pour tout $y \in \mathrm{Im}\,A$, $\|z - y\| \geq \|(I - \Pi)z\|$. Or il existe un (et un seul) point $y \in \mathrm{Im}(A)$ qui atteint cette borne : il s'agit du point $y = \Pi z$. En effet :

$$\|z - \Pi z\|^2 = \|\Pi(z - \Pi z))\|^2 + \|(I - \Pi)z\|^2 = \|(I - \Pi)z\|^2\,,$$

en utilisant à nouveau le fait que $\Pi^2 = \Pi$. Cela signifie que le point $y = \Pi z$ est le point de $\mathrm{Im}(A)$ qui est le plus proche de z. □

Dans le cadre du Théorème 1.3, on notera que la matrice $A^\# := (A^TA)^{-1}A^T$ vérifie la propriété suivante

$$A^\#A = I$$

et cette matrice $A^\#$ est dite pseudo-inverse à gauche.

1.3. Exercices

Algèbre linéaire

Exercice 1.1. Montrer que $\ker A = \ker(A^T A)$.

Exercice 1.2. Soient deux matrices symétriques A et B Montrer que si $A \succeq B$ alors $a_{\ell,\ell} \geq b_{\ell,\ell}$ pour tout $\ell = 1, \cdots, d$ et donc $\mathrm{trace}(A) \geq \mathrm{trace}(B)$.

Exercice 1.3. Soit x un vecteur-(colonne) de $\mathbb{R}^m$. Démontrer que $x^T.x = \mathrm{trace}(x.x^T)$.

Exercice 1.4. Soit une matrice A carrée symétrique semi-définie positive. Démontrer qu'il existe une matrice carrée de même taille, notée Γ, telle que, $A = \Gamma^T.\Gamma$. En déduire que $x^T A x = \|\Gamma x\|^2$.

Exercice 1.5. Démontrer l'identité $\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$.

Exercice 1.6. Soit Π_A le projecteur sur $\mathrm{Im}(A)$ où A est une matrice $m \times d$. Montrer l'identité de Pythagore :

$$\|z\|^2 = \|\Pi_A z\|^2 + \|(I - \Pi_A)z\|^2 .$$

Exercice 1.7. Dans le cadre du Théorème 1.3, si $m > s$, montrer que $AA^{\#}$ ne peut être égal à l'identité. Si $m = d$, que se passe t-il pour le projecteur ?

Optimisation

Exercice 1.8. Soit $f : \mathbb{R}^d \to \mathbb{R}$. Calculer $\nabla f(x)$ et $\mathrm{Hess}(f)$ dans les cas suivants :

- $f(x) = \frac{1}{2}\|x\|^2$.
- $f(x) = \frac{1}{2}\|Ax - b\|^2$.

Ces fonctions sont-elles convexes ? Caractériser $\arg\min f$.

2. Probabilités

2.1. Généralités

Une *expérience aléatoire* est une expérience pouvant conduire à plusieurs résultats possibles. Formellement, une expérience aléatoire se décrit par la donnée de l'ensemble Ω des résultats possibles. L'ensemble Ω est appelé l'*univers* ou l'*espace des états*.

Traditionnellement, un résultat possible de l'expérience est noté ω. C'est un élément de l'univers Ω. Un tel élément $\omega \in \Omega$ est parfois appelé une *épreuve* ou une *issue*.

Un *événement aléatoire* est un événement dont la réalisation dépend du résultat de l'expérience. Formellement, un événement aléatoire se décrit comme un sous-ensemble de Ω.

Pour une issue donnée $\omega \in \Omega$, on dit qu'un événement A est *réalisé* si $\omega \in A$.
L'espace d'état Ω est aussi appelé l'*événement certain* : il est réalisé quelle que soit l'issue.
L'ensemble vide $\emptyset$ est aussi appelé l'*événement impossible* : il n'est jamais réalisé.

Expérience aléatoire	*Univers*	*Exemple d'événement*
Jet de dé	$\Omega = \{1, 2, 3, 4, 5, 6\}$	$A = \{2, 4, 6\}$ « le résultat est pair »
Deux lancers consécutifs d'une pièce	$\Omega = \{PP, PF, FP, FF\}$ où P = pile, F = face	$A = \{PP, FF\}$ « on obtient deux faces identiques »
Durée de fonctionnement sans panne d'une machine	$\Omega = [0, +\infty)$	$A = [x, +\infty)$ « La machine fonctionne pendant au moins x unités de temps »
Valeur d'un signal sur un intervalle de temps	Ω =ensemble des fonctions continues sur $[t_0, t_1]$	$A = \{\omega \in \Omega : \sup_t \lvert\omega(t)\rvert \leq \alpha\}$ « l'amplitude du signal n'excède pas α »

TABLE 2.1. – Exemples d'expériences aléatoires.

Définition 2.1. Une mesure de probabilité $\mathbb{P}$ sur un ensemble Ω au plus dénombrable est une application qui à un sous-ensemble A de l'univers Ω associe un nombre $\mathbb{P}(A)$, tel que :

— $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$,

— pour toute famille $(A_j,\ j \in \mathbb{N}^*)$ de parties deux à deux disjointes de Ω,

$$\mathbb{P}(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{+\infty} \mathbb{P}(A_j). \tag{2.1}$$

Proposition 2.2. *Soit $\mathbb{P}$ une probabilité sur Ω. Soient A, B, $(A_n)_{n\in\mathbb{N}^*}$ des sous-ensembles de Ω. Alors,*

a) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, *où A^c est le complémentaire de A.*

b) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

c) *Si $(A_n)_{n\in\mathbb{N}^*}$ est une partition de Ω, c'est à dire deux à deux disjoints, et dont l'union est Ω, alors*

$$\mathbb{P}(B) = \sum_{n=1}^{\infty} \mathbb{P}(A_n \cap B) \ .$$

d) *Pour une famille quelconque $(A_n)_{n\in\mathbb{N}^*}$ dans $\mathcal{F}$, on a la* borne de l'union *:*

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n) \ .$$

2.1.1. Probabilité conditionnelle : définition

De façon informelle, la probabilité d'un événement vise à quantifier l'occurence de cet événement. La probabilité conditionnelle d'un événement A sachant un événement B vise à quantifier l'occurence de A lorsque l'on sait que B s'est produit. D'un point de vue plus formel, on a la définition suivante.

Soit $\mathbb{P}$ une probabilité sur Ω.

Définition 2.3. Pour tous événements A, B tels que $\mathbb{P}(B) \neq 0$, on définit la probabilité conditionnelle de A sachant B, et on note $\mathbb{P}(A \,|\, B)$, la quantité :

$$\mathbb{P}(A \,|\, B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \ .$$

Si on associe probabilité et « poids », la probabilité d'un ensemble étant son poids relatif par rapport à celui de l'ensemble total, la probabilité conditionnelle de A sachant B est le poids de la trace de A sur B *relativement* au poids total de B.

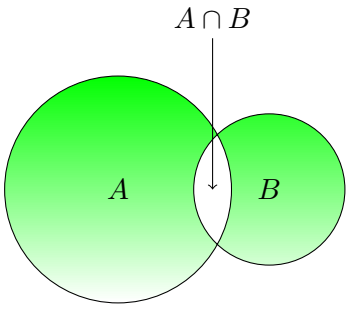


FIGURE 2.1. – Interprétation graphique du conditionnement.

Considérons le cas où $\mathbb{P}$ est la probabilité uniforme sur un ensemble Ω fini, c'est-à-dire $\mathbb{P}(A) = |\,A\,|\,/\,|\,\Omega\,|$. On a alors $\mathbb{P}(A \,|\, B) = |\,A \cap B\,|\,/\,|\,B\,|$. Cette expression justifie la remarque suivante :

$\mathbb{P}(A \mid B)$ peut être interprétée comme la probabilité de l'événement $A \cap B$ dans ce nouvel univers qu'est B.

Exemple 2.4. Considérons le lancer d'un dé : $\mathbb{P}$ est la probabilité uniforme sur $\Omega = \{1, 2, \ldots, 6\}$. Calculer la probabilité d'obtenir « 6 » sachant que le résultat est pair.

Exemple 2.5. On dispose de trois pièces de monnaie : l'une est bien équilibrée, l'une comporte deux côtés `pile`, l'autre deux côtés `face`. On choisit une pièce au hasard. Evaluons la probabilité de tomber sur `pile`.

Désignons par E, $2P$ et $2F$ les événements « la pièce bien équilibrée est choisie », « la pièce comportant deux côtés `pile` est choisie », etc. D'après la propriété ci-dessus,

$$\begin{aligned} \mathbb{P}(\texttt{pile}) &= \mathbb{P}(\texttt{pile} \mid E)\mathbb{P}(E) + \mathbb{P}(\texttt{pile} \mid 2P)\mathbb{P}(2P) + \mathbb{P}(\texttt{pile} \mid 2F)\mathbb{P}(2F) \\ &= \frac{1}{2} \times \frac{1}{3} + 1 \times \frac{1}{3} + 0 \times \frac{1}{3} \;=\; \frac{1}{2}\,. \end{aligned}$$

La seconde propriété est connue sous le nom de *formule de Bayes*. La preuve est immédiate.

Proposition 2.6. *Soient A, B deux événements tels que $\mathbb{P}(A) \neq 0$ et $\mathbb{P}(B) \neq 0$. Alors,*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}\,.$$

La formule de Bayes permet typiquement d'évaluer des probabilités du type :

$$\mathbb{P}(\text{une action « a » a été effectuée} \mid \text{le résultat « r » a été observé})$$

lorsqu'on connait le modèle $\mathbb{P}(\text{le résultat « r » est observé} \mid \text{l'action « a » est effectuée})$.

Exemple 2.7. Reprenons l'exemple précédent des trois pièces. Sachant qu'on obtient le résultat `pile`, quelle est la probabilité que la pièce à deux côtés pile ait été choisie ? La réponse est donnée par la formule de Bayes :

$$\mathbb{P}(2P \mid \texttt{pile}) = \frac{\mathbb{P}(\texttt{pile} \mid 2P)\,\mathbb{P}(2P)}{\mathbb{P}(\texttt{pile})} = \frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}\,.$$

2.1.2. Événements indépendants

Dans l'exemple précédent, l'événement $B =$« `pile` est le résultat » apporte une information sur la probabilité que l'événement $A =$« la pièce à deux côtés `pile` a été choisie ». Avant l'expérience qui a vu B se réaliser, la probabilité de A valait $\frac{1}{2}$. Après l'expérience, elle vaut $\frac{2}{3}$. Le fait que B soit réalisé ne dit pas si A est ou non réalisé, mais par contre, il change notre croyance en A. A l'inverse, il existe des événements A, B tels que la réalisation de B n'apporte aucune information sur A. De tels événements sont dits *indépendants*. Voici une définition plus formelle.

Définition 2.8. Deux événements A, B sont dits *indépendants*, noté $A \perp\!\!\!\perp B$, si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B) \ .$$

Si $\mathbb{P}(B) \neq 0$, la définition revient bien à $\mathbb{P}(A \,|\, B) = \mathbb{P}(A)$: la réalisation de B ne modifie pas la croyance en A.

Remarque 2.1. a) Les propriétés suivantes sont équivalentes : $A \perp\!\!\!\perp B$, $B \perp\!\!\!\perp A$, $A \perp\!\!\!\perp B^c$, $A^c \perp\!\!\!\perp B$, $A^c \perp\!\!\!\perp B^c$.
b) Si $\mathbb{P}(B) = 0$ ou $\mathbb{P}(B) = 1$, alors A et B sont indépendants quel que soit A.

Définition 2.9. Soit I un ensemble quelconque. Une famille $(A_i)_{i\in I}$ d'événements est dite indépendante si pour tout sous-ensemble **fini** $J \subset I$, on a :

$$\mathbb{P}\left(\bigcap_{j\in J} A_j\right) = \prod_{j\in J} \mathbb{P}(A_j) \ .$$

Illustrons la formule ci-dessus lorsque la famille contient trois événements A, B, C. Les événements A, B, C sont indépendants si

$$\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B),\ \mathbb{P}(A\cap C) = \mathbb{P}(A)\mathbb{P}(C),\ \mathbb{P}(B\cap C) = \mathbb{P}(B)\mathbb{P}(C),$$
$$\text{et } \mathbb{P}(A\cap B\cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) \ .$$

Il est important de souligner que la première ligne d'équations ci-dessus n'implique pas la deuxième : ce n'est pas parce que des événements sont deux à deux indépendants qu'ils forment une famille indépendante.

Exemple 2.10. On lance deux dés. On note A l'événement « le premier lancer est pair », B l'événement « le second lancer est pair » et C l'événement « la somme des deux lancers est paire ». Les événements A, B, C ne sont pas indépendants car $\mathbb{P}(A\cap B\cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. Pourtant, A, B sont indépendants, A, C sont indépendants, B, C sont indépendants.

2.2. Variables aléatoires discrètes

2.2.1. Loi, indépendance

Une variable aléatoire discrète (v.a.d.) est une grandeur à valeur dans $\mathbb{N}$ qui dépend du résultat de l'expérience. C'est donc une *fonction* de l'issue ω (en ce sens, la terminologie de *variable* est assez malencontreuse). On la notera souvent :

$$\begin{array}{rccc} X \ : & \Omega & \to & \mathbb{N} \\ & \omega & \mapsto & X(\omega) \ . \end{array}$$

Exemple 2.11. Considérons un lancer de n dés. Une issue ω est un n-uplet sur $\Omega = \{1, \cdots, 6\}^n$. On peut par exemple définir la variable aléatoire $X(\omega)$ qui est égale au nombre de « 6 » obtenus : c'est bien une fonction de ω.

Notons que, au delà de $\mathbb{N}$, tout ce suit s'étend au cas de variables aléatoires dans un espace fini ou dénombrable. Nous nous contentons de $\mathbb{N}$ pour rester aussi concrets que possible.

En probabilités, on s'intéresse plus particulièrement à évaluer la probabilité d'événements de la forme « la variable X vaut x » ou, plus généralement,

$$\text{« la variable } X \text{ appartient à l'ensemble } H \text{ »} = \{\omega \in \Omega \,:\, X(\omega) \in H\}$$

pour un sous-ensemble $H \subset E$ arbitraire. Nous utiliserons souvent la notation $\{X \in H\}$ pour désigner l'événement $\{\omega \in \Omega \,:\, X(\omega) \in H\}$. La probabilité $\mathbb{P}(\{X \in H\})$ de l'événement $\{X \in H\}$ est souvent notée en omettant les accollades, soit $\mathbb{P}(X \in H)$.

Définition 2.12. Soit X une v.a.d. On appelle *loi* de X l'ensemble des coefficients

$$\mathbb{P}(X = k) \quad \text{pour tout } k \in \mathbb{N}.$$

Les lois discrètes les plus importantes sont données dans la Table 2.2. A partir de ces coefficients, on peut calculer toutes les probabilités du type $\mathbb{P}(X \in H)$, par exemple :

$$\mathbb{P}(X \leq 2) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2).$$

Si X et Y sont deux v.a. sur $\mathbb{N}$, on définit la *loi jointe* du couple (X, Y) comme l'ensemble des coefficients :

$$\mathbb{P}(X = k, Y = \ell)$$

pour tous entiers k, ℓ. Le nombre ci-dessus est à comprendre comme la probabilité de l'ensemble des ω tels que $X(\omega) = k$ ET $Y(\omega) = \ell$. Les lois de X et Y (c'est à dire les coefficients $\mathbb{P}(X = k)$ et $\mathbb{P}(Y = \ell)$ respectivement) sont appelées les *lois marginales* de X et de Y.

Définition 2.13. Les v.a. X et Y sont dites *indépendantes* si pour tous k, ℓ,

$$\mathbb{P}(X = k, Y = \ell) = \mathbb{P}(X = k)\mathbb{P}(Y = \ell),$$

c'est à dire si les événements $\{X = k\}$ et $\{Y = \ell\}$ sont indépendants.

Dans ce cas, on a en particulier, pour tous ensembles H et G,

$$\mathbb{P}(X \in G, Y \in H) = \mathbb{P}(X \in G)\mathbb{P}(Y \in H)\,.$$

On s'intéresse souvent à des fonctions de variables aléatoires, du type $f(X)$, où f est une fonction. Ici, $f(X)$ désigne la v.a. qui à tout $\omega \in \Omega$ associe $f(X(\omega))$ (par exemple, X^2 pour $f(x) = x^2$).

Proposition 2.14. *Si X et Y sont indépendantes, alors pour toutes fonctions f et g, pour tous H, G,*

$$\mathbb{P}(f(X) \in H, g(Y) \in G) = \mathbb{P}(f(X) \in H)\mathbb{P}(g(Y) \in G)\,,$$

ce qui signifie que les variables aléatoires $f(X)$ et $g(Y)$ sont elles-mêmes indépendantes.

Variable	Support	$\mathbb{P}(X=k)$	$\mathbb{E}(X)$	$\mathrm{Var}(X)$
Bernoulli $\mathcal{B}(p)$	$\{0,1\}$	$p^k(1-p)^{1-k}$	p	$p(1-p)$
Uniforme $\mathcal{U}(N)$	$\{0,1,\dots,N\}$	$\frac{1}{N}$	$*$	$*$
Binomiale $\mathcal{B}(p,n)$	$\{0,1,\dots,n\}$	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$
Poisson $\mathcal{P}(\lambda)$	$\mathbb{N}$	$\frac{\lambda^k}{k!}e^{-\lambda}$	λ	λ
Géométrique $\mathcal{G}(p)$	$\mathbb{N}^*$	$p(1-p)^{k-1}$	$1/p$	$*$

TABLE 2.2. – Principales variables discrètes ($*$= faites le calcul si vous le souhaitez).

2.2.2. Espérance, variance, moments

Définition 2.15. L'*espérance* de X est définie par

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k\,\mathbb{P}(X=k)\,. \tag{2.2}$$

Le résultat suivant, appelé *Théorème de transfert* permet d'exprimer l'espérance d'une v.a. $g(X)$ (c'est à dire d'une fonction de X) directement en fonction de la loi de X.

Proposition 2.16. *Soit g est une fonction à valeurs réelles, on a :*

$$\mathbb{E}(g(X)) = \sum_{k=0}^{\infty} g(k)\,\mathbb{P}(X=k)\,.$$

Le moment d'ordre p est $\mathbb{E}(X^p)$. On va tout particulièrement s'intéresser au moment d'ordre deux, $\mathbb{E}(X^2)$, et à la variance définie par

$$\mathrm{Var}(X) = \mathbb{E}((X-\mathbb{E}(X))^2)\,.$$

On a aussi que $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. L'écart-type est défini par :

$$\sigma_X = \sqrt{\mathrm{Var}(X)}\,.$$

Rappelons les exemples principaux, dans la Table 2.2. Si g est une fonction quelconque à valeurs réelles, on a :

$$\mathbb{E}(g(X,Y)) = \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} g(k,\ell)\mathbb{P}(X=k,Y=\ell)\,.$$

Les v.a. X et Y sont dites *indépendantes* (noté $X \perp\!\!\!\perp Y$) si

$$\mathbb{P}(X=k,Y=\ell) = \mathbb{P}(X=k)\mathbb{P}(Y=\ell)$$

pour tout couple (k,ℓ).

Proposition 2.17. *Si X et Y sont indépendantes, la propriété suivante est vraie pour toutes fonctions h et g :*

$$\mathbb{E}(h(X)g(Y)) = \mathbb{E}(h(X))\mathbb{E}(g(Y))\,. \tag{2.3}$$

2.3. Variables aléatoires à densité

2.3.1. Définitions

Certaines variables aléatoires ne sont pas à valeurs dans $\mathbb{N}$ ou dans un espace discret, mais peuvent prendre *a priori* n'importe quelle valeur dans l'ensemble des réels, ou plus généralement dans un autre ensemble non-dénombrable. On parle aussi parfois de variables aléatoires "continues", même si cette terminologie est impropre.

Exemple 2.18. $X(\omega)$ = la durée de vie sans panne d'une machine, $X(\omega)$ peut prendre *a priori* n'importe quelle valeur positive.

Exemple 2.19. $X(\omega)$ = la position à l'instant t d'une particule se déplaçant aléatoirement dans l'espace. Dans ce cas, $X(\omega)$ est une valeur de $\mathbb{R}^3$.

Dans les exemples précédents, il se trouve que, quel que soit x, $\mathbb{P}(X = x) = 0$. Par exemple, la machine tombera en panne, mais la probabilité que cela se produise exactement après 1 jour, 3 heures, 15 minutes, 0 secondes, est nulle. On ne peut donc plus caractériser la loi de la v.a. X par les coefficients $\mathbb{P}(X = x)$ pour x variant dans l'ensemble des valeurs possibles prises par X. Dans l'exemple de la machine, on peut exprimer la probabilité que la machine tombe en panne avant un instant x, par :

$$\mathbb{P}(X \leq x)\,.$$

L'ensemble de ces coefficients caractérise la loi de la v.a. X.

Définition 2.20. La fonction de répartition d'une variable aléatoire à valeurs réelle $X : \Omega \to \mathbb{R}$ est définie par :

$$F_X(x) = \mathbb{P}(X \leq x)\,.$$

On dit qu'une v.a. $X : \Omega \to \mathbb{R}$ admet une *densité de probabilité*, s'il existe une fonction $p_X : \mathbb{R} \to [0, +\infty)$ telle que :

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} p_X(t)dt\,.$$

Cela implique en particulier que la fonction de répartition F_X est continue. La fonction p_X est appelée la densité de probabilité de X (ou juste, la densité, pour faire court). En faisant tendre $x \to +\infty$ dans la définition précédente, le membre de gauche converge vers $\mathbb{P}(X \leq +\infty)$ qui vaut 1. On en déduit la propriété suivante d'une densité de probabilité :

$$\int p_X(x)dx = 1\,.$$

Ainsi, la fonction de répartition est l'intégrale de la densité et, réciproquement, la densité de probabilité est la dérivée de la fonction de répartition :

$$p_X(x) = F'_X(x)$$

Et inversement, lorsque l'on connaît la densité x d'une v.a., on peut facilement calculer sa fonction de répartition en intégrant.

Exemple 2.21. Afin de modéliser la durée de panne d'une machine, ou encore la survie d'un patient après un diagnostic, on utilise fréquemment l'hypothèse que X suit une *loi exponentielle*. On entend par là que X est une variable à densité, dont la densité est donnée par :

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

où $\lambda > 0$ est un paramètre. En calculant la primitive de p_X, on obtient que $P_X(x) = 1 - e^{-\lambda x}$, pour tout $x \geq 0$.

Par la formule des probabilités totales, pour tout $a < b$, on a $\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b)$, donc :

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b p_X(x)dx\,. \tag{2.4}$$

En particulier, en faisant tendre b vers a, on obtient que, si X admet une densité,

$$\text{Pour tout } a \in \mathbb{R},\ \mathbb{P}(X = a) = 0\,.$$

Autrement dit, la probabilité que $X(\omega)$ soit exactement égale à un nombre, est toujours nulle.

Remarque 2.2. Toutes les v.a. n'admettent pas forcément de densité. Par exemple, une variable discrète, comme par exemple, une variable de Bernoulli sur $\{0, 1\}$, satisfait $\mathbb{P}(X = 1) > 0$, ce qui serait en contradiction avec l'existence d'une densité. Dans ce cours, nous ne considérons que deux cas : ou bien X admet une densité, ou bien c'est une variable aléatoire discrète. Il existe des cas "hybrides", où X n'est pas discrète, et pour autant n'admet pas de densité [1] Pour établir une théorie des probabilités complètes, permettant de couvrir ces cas, il faut se tourner vers d'autres outils mathématiques, en l'occurrence la théorie de la mesure. Il n'en sera pas question ici.

L'égalité (2.4) se généralise de la manière suivante.

Proposition 2.22. *Si X est un v.a. réelle admettant une densité p_X, et si $H \subset \mathbb{R}$ est un ensemble, on a :*

$$\mathbb{P}(X \in H) = \int_H p_X(x)dx\,, \tag{2.5}$$

c'est à dire que la probabilité que $X \in H$ s'obtient en intégrant la densité sur l'ensemble H uniquement.

2.3.2. Espérance

La définition de l'espérance comme dans le cas des v.a. discrètes (2.2) n'a plus de sens pour les v.a. à densité, puisque $\mathbb{P}(X = k)$ est nul pour tout k.

1. Considérer par exemple la variable $X = \min(1, Z)$ où Z suit une loi exponentielle. La variable X n'est pas discrète, car elle peut *a priori* prendre n'importe quelle valeur dans $[0, 1]$, et pourtant, elle n'admet pas de densité, car $\mathbb{P}(X = 1) = \mathbb{P}(Z \geq 1) = 1 - F_Z(1) = e^{-\lambda}$. Puisque $\mathbb{P}(X = 1) > 0$, il est clair que X ne peut admettre de densité.

Variable	Support	Densité	Espérance	Variance
Uniforme $\mathcal{U}([a,b])$	$[a,b]$	$(b-a)^{-1}\mathbf{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{2}$
Exponentielle $\mathcal{E}(\lambda)$	$[0,+\infty)$	$\lambda e^{-\lambda x}\mathbf{1}_{[0,\infty)}(x)$	λ^{-1}	λ^{-2}
Gaussienne $\mathcal{N}(m,\sigma^2)$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-m)^2}{2\sigma^2}}$	m	σ^2
Chi-deux $\chi^2(k)$	$[0,+\infty)$	$\frac{1}{2^{k/2}\Gamma(k/2)}\ x^{k/2-1}e^{-x/2}$	k	$2k$
Student $\mathcal{T}(k)$	$\mathbb{R}$	$\frac{1}{\sqrt{k\pi}}\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}$	0 (si $k>1$)	$\frac{k}{k-2}$ (si $k>2$)

TABLE 2.3. – Principales variables à densité. La loi $\mathcal{N}(0,1)$ s'appelle la loi gaussienne (ou normale) centrée réduite. La lettre Γ représente la fonction Gamma d'Euler. La loi $\chi^2(k)$, ou loi du chi-deux à k degrés de liberté ($k \geq 1$), est la loi d'une somme de la somme $X_1^2 + \cdots + X_k^2$, où $X_1, \ldots, X_k$ sont iid gaussiennes centrées réduites. La loi $\mathcal{T}(k)$, ou loi de Student à k degrés de liberté, est la loi du rapport $\frac{Z}{\sqrt{U/k}}$, où $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2(k)$ et U, Z sont indépendantes.

Définition 2.23. L'espérance d'une v.a.r. X de densité p_X est définie par :

$$\mathbb{E}(X) = \int x p_X(x) dx\,,$$

et plus généralement, si g est une fonction réelle,

$$\mathbb{E}(g(X)) = \int g(x) p_X(x) dx\,.$$

Autrement dit, l'espérance est le barycentre de la densité. On peut comprendre $\mathbb{E}(X)$ comme la valeur "moyenne" prise par $X(\omega)$, sa "tendance centrale". La variance est, comme dans le cas discret, $\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$ et l'écart-type est la racine carrée.

2.3.3. Vecteurs aléatoires et indépendance

Deux v.a.r. X et Y définissent un vecteur aléatoire (X,Y) sur $\mathbb{R}^2$. La fonction de répartition $F_{X,Y}$ de ce vecteur est définie par :

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

De même, on appelle *densité jointe* du couple (X,Y), si elle existe, la fonction positive $p_{X,Y}$ telle que

$$\forall x, y, \quad F_{X,Y}(x,y) = \int_\infty^x \int_\infty^y p_{X,Y}(s,t) ds dt,$$

ou, autrement dit,

$$p_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}\,.$$

Les densités respectives de X et Y, appelées *densités marginales* sont alors liées à la densité jointe par :

$$p_X(x) = \int p_{X,Y}(x,y) dy \quad \text{et} \quad p_Y(y) = \int p_{X,Y}(x,y) dx\,.$$

Si $g(x, y)$ est une fonction réelle de deux variables, on a :

$$\mathbb{E}(g(X, Y)) = \int\int g(x, y)p_{X,Y}(x, y)dxdy\,.$$

En particulier, dans le cas où $g(x, y) = \mathbf{1}_H(x, y)$ est l'indicatrice d'un ensemble $H \subset \mathbb{R}^2$ quelconque, l'espérance du membre de gauche se ramène à une probabilité, et on peut écrire :

$$\mathbb{P}((X, Y) \in H) = \int\int p_{X,Y}(x, y)\mathbf{1}_H(x, y)dxdy$$

c'est à dire que la probabilité qu'un vecteur aléatoire appartienne à une certaine région $H \subset \mathbb{R}^2$ est l'intégrale de la densité, sur cette région H.

Dans le cas particulier où $p_{X,Y}(x, y) = p_X(x)p_Y(y)$, on dit que les deux v.a.r. X et Y sont *indépendantes*. Dans ce cas, l'égalité (2.3) est satisfaite.

Toutes ces notions et ces résultats se généralisent évidemment au cas de non plus deux, mais d variables aléatoires.

2.3.4. Inégalités

Soient X, Y des v.a.r. On rappelle l'inégalité de Cauchy-Schwarz.

Proposition 2.24. *On a :*

$$\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}\,.$$

De plus, le cas d'égalité $\mathbb{E}(XY) = \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}$ *signifie soit que* Y *est la variable nulle, soit que* $\exists\lambda \geq 0,\ X = \lambda Y$.

Soit $\epsilon > 0$. L'inégalité de Bienaymé-Tchebychev, que nous ne redémontrons pas, est donnée par :

$$\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}(X^2)}{\epsilon^2}\,. \tag{2.6}$$

2.3.5. Quantiles

Soit X une variable aléatoire à densité. Pour tout α entre 0 et 1, on appelle *quantile de niveau* α, le nombre $q_\alpha \in \mathbb{R}$ tel que :

$$\mathbb{P}(X \leq q_\alpha) = \alpha\,.$$

On suppose ici implicitement que ce nombre existe et est défini de façon unique, ce qui est vrai dans les tous cas d'usage. Autrement dit, si F_X désigne la fonction de répartition de X, cela se lit : $F_X(q_\alpha) = \alpha$, soit :

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

où F_X^{-1} est l'inverse de la fonction de répartition.

Le quantile de niveau 0.5 s'appelle la médiane. Le quantile de niveau 0.25 s'appelle le premier quartile. Le quantile de niveau 0.75 s'appelle le troisième quartile. Le quantile de niveau 0.1 s'appelle le premier décile. Le quantile de niveau 0.2 s'appelle le deuxième décile. Et ainsi de suite.

2.3.6. Covariance, corrélation

Le coefficient suivant est appelé la *covariance* de X et Y :

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))\,.$$

On introduit le *coefficient de corrélation* entre X et Y, défini par :

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}\,.$$

Le coefficient de corrélation est une version renormalisée de la covariance, comme l'indique la proposition suivante :

Proposition 2.25. *On a :*

$$-1 \leq \rho_{X,Y} \leq 1\,.$$

On suppose que les v.a. X, Y ne sont pas constantes. Alors :

- *si $\rho_{XY} = 1$, alors il existe $a > 0$ et $b \in \mathbb{R}$, tels que $Y = aX + b$.*
- *si $\rho_{XY} = -1$, alors il existe $a < 0$ et $b \in \mathbb{R}$, tels que $Y = aX + b$.*

Démonstration. C'est une conséquence immédiate de l'inégalité de Cauchy-Schwarz. □

On dit que deux v.a. sont *décorrélées* si $\mathrm{Cov}(X,Y) = 0$, soit de manière équivalente $\rho_{X,Y} = 0$. D'après l'équation (2.3), l'indépendance de deux v.a. implique la décorrélation (et la réciproque est fausse en général, hormis dans le cas notable des vecteurs gaussiens que nous verrons plus bas). Donc, si deux variables aléatoires sont corrélées, cela implique qu'elles sont dépendantes. Les deux cas extrêmes sont donnés par $\rho_{XY} = \pm 1$. Dans ce cas, la dépendance est totale car on peut même affirmer que Y s'écrit comme une fonction (affine) de X. Cela signifie en particulier que la donnée de X *détermine* entièrement la valeur de Y, au travers d'une relation affine. En ce sens, le coefficient de corrélation ρ_{XY} peut être interprété comme une mesure de dépendance (affine) entre X et Y.

La corrélation $\rho_{X,Y}$ peut être positive ou négative. Une corrélation positive tend à indiquer que de grandes valeurs de X vont de pair avec de grandes valeurs de Y. Une corrélation négative tend à indiquer que de grandes valeurs de X vont de pair avec de faibles valeurs de Y.

2.3.7. Matrice de covariance

La donnée de d v.a.r. $X_1, X_2, \ldots, X_d$ défini un vecteur aléatoire, que nous supposerons dorénavant être un vecteur-colonne $d \times 1$, noté $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix} . \tag{2.7}$$

On notera sa densité jointe par $p_{\boldsymbol{X}}(x_1, \ldots, x_d)$. On définira l'espérance d'un vecteur aléatoire comme étant le vecteur des espérances :

$$\mathbb{E}(\boldsymbol{X}) = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_d) \end{pmatrix} .$$

De même, on définira la *matrice de covariance* $\mathrm{Cov}(\boldsymbol{X})$ du vecteur aléatoire $\boldsymbol{X}$ comme étant la matrice $d \times d$ formé par l'ensemble des coefficients $\mathrm{Cov}(X_i, X_j)$ pour i, j décrivant $\{1, \ldots, d\}$, soit :

$$\mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & & & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Cov}(X_d, X_d) \end{pmatrix} .$$

On rappelle que :

- La diagonale de $\mathrm{Cov}(\boldsymbol{X})$ correspond aux variances $\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_d)$, car $\mathrm{Cov}(X_1, X_1) = \mathrm{Var}(X_1)$
- $\mathrm{Cov}(\boldsymbol{X})$ est une matrice symétrique car $\mathrm{Cov}(X_1, X_2) = \mathrm{Cov}(X_2, X_1)$.
- $\mathrm{Cov}(\boldsymbol{X})$ est une matrice semi-définie positive, en ce sens que pour tout vecteur-colonne $x \in \mathbb{R}^d$, $x^T \mathrm{Cov}(X) x \geq 0$.

2.3.8. Vecteurs gaussiens

Soit $\boldsymbol{X} : \Omega \to \mathbb{R}^d$ un vecteur-colonne, de la forme (2.7). On dit que $\boldsymbol{X}$ est un vecteur gaussien si, pour tout vecteur $x \in \mathbb{R}^d$, le produit scalaire $x^T X$ est une variable gaussienne (c'est à dire, qui suit la loi $\mathcal{N}(m, \sigma^2)$ pour un certain m et un certain σ^2). Si $\boldsymbol{X}$ est un vecteur gaussien, alors $X_1, \ldots, X_d$ sont des v.a. gaussiennes. Attention, la réciproque n'est pas toujours vraie, mais elle est vraie au moins dans le cas donné par l'exemple suivant.

Exemple 2.26. Si $X_1, \ldots, X_d$ sont des v.a. gaussiennes indépendantes, alors le vecteur (2.7) est un vecteur gaussien. En outre, sa matrice de covariance est diagonale.

La première propriété fondamentale des vecteurs gaussiens est la suivante :

Toute transformation affine d'un vecteur gaussien est un vecteur gaussien.

Autrement dit, si $\boldsymbol{X}$ est un vecteur gaussien, alors tout vecteur aléatoire de la forme $A\boldsymbol{X}+b$ est un vecteur gaussien (où A est une matrice, et b un vecteur, tous deux déterministes).

La seconde propriété fondamentale des vecteurs gaussiens est la suivante :

Si la matrice de covariance d'un vecteur gaussien est diagonale, alors les composantes de ce vecteur sont indépendantes.

Par exemple, on sait qu'en général, si X et Y sont deux v.a.r. décorrélées, elles ne sont pas nécessairement indépendantes. Par contre, si on sait en outre que $(X, Y)^T$ est un vecteur gaussien, alors décorrélation vaut indépendance.

On utilise la notation

$$\boldsymbol{X} \sim \mathcal{N}(m, \Sigma)$$

pour écrire que $\boldsymbol{X}$ est un vecteur gaussien d'espérance m et de matrice de covariance Σ. Dans le cas où Σ est inversible, alors $\boldsymbol{X}$ admet une densité, qui est donnée par la formule suivante, pour tout $x = (x_1, \dots, x_d)^T$:

$$p_{\boldsymbol{X}}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} e^{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)}\,.$$

2.4. Loi des grands nombres

2.4.1. Résultat

Théorème 2.27. *Soit une suite de v.a.r. $(X_n)_{n\in\mathbb{N}^*}$ i.i.d. (indépendantes et identiquement distribuées), et d'espérance m. Alors,*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = m\right) = 1\,.$$

On dit aussi que la moyenne empirique $\frac{1}{n}\sum_{i=1}^n X_i$ converge *presque sûrement* (c'est à dire avec probabilité un), vers l'espérance, ce que l'on note :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p.s.} \mathbb{E}(X_1)\,.$$

Dans le membre de droite de la limite ci-dessus, on a écrit $\mathbb{E}(X_1)$, mais on aurait pu écrire $\mathbb{E}(X_i)$ pour n'importe quel i, puisque toutes ces espérances sont égales à la même valeur m : cela vient du fait que les X_i sont supposés identiquement distribués.

Exemple 2.28. Soit (X_n) une suite i.i.d. de variables de Bernoulli sur $\{0, 1\}$, de paramètre $0 < p < 1$. Pour chaque issue possible ω de l'expérience aléatoire, $(X_1(\omega), X_2(\omega), \dots)$ est une suite de 0 et de 1. Typiquement, pour un certain ω, on pourrait rencontrer une suite du type

$$(1, 0, 1, 1, 0, 0, 1, 0, \dots)\,.$$

La loi des grands nombres établit que, avec probabilité un, l'issue ω de l'expérience aléatoire sera telle que la moyenne de Cesaro de cette suite converge vers $\mathbb{E}(X_1)$, qui vaut p dans le cas présent. Autrement dit, la fréquence empirique des "1" converge vers la probabilité p d'obtenir "1". Notons que certaines issues particulières ω ne vérifient pas cette propriété : par exemple, une issue possible de l'expérience est :

$$(1, 1, 1, 1, 1, 1, \dots)$$

et pour cette issue, la moyenne empirique de X_i converge vers 1, et non vers p. La loi des grands nombres établit que ce type de cas se produit avec une probabilité nulle.

2.4.2. Histogramme

Voyons une application de la loi des grands nombres en statistique.

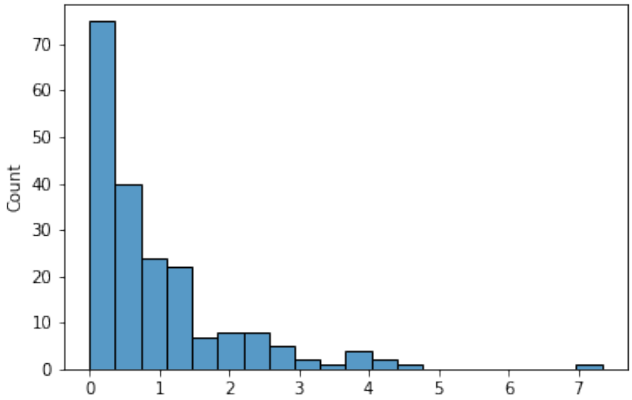


FIGURE 2.2. – Histogramme associé à un tirage aléatoire de n =200 points, selon un loi exponentielle de paramètre 1. On a choisi $M = 20$ bins.

Admettons que l'on observe n points dans $\mathbb{R}$, que nous interprétons comme des réalisations de variables aléatoires i.i.d. $X_1, \dots, X_n$, toutes de même densité p_X. On appelle cela un n-échantillon. Pour simplifier, on suppose que les variables prennent leur valeurs dans un intervalle $[a, b]$. On décide de découper l'intervalle $[a, b]$ en M segments $I_1, \dots, I_M$ de même longueur $(b - a)/M$, c'est à dire :

$$I_1 = [a, a + \tfrac{b-a}{M}),\ I_2 = [a + \tfrac{b-a}{M}, a + 2\tfrac{b-a}{M}), \dots, I_M = [a + \tfrac{(M-1)(b-a)}{M}, b]\,.$$

L'histogramme du n-échantillon est un graphique en "barres" : il y a M barres (ou *bins*), et la hauteur de chaque barre est égale au nombre d'échantillons X_i appartenant au segment I_i. Formellement, l'histogramme est une fonction constante par morceaux, dont la valeur sur le segment I_m (pour $m = 1, \dots, M$) est :

$$C_{n,m} = \mathrm{Card}\{i = 1, \dots, n\ :\ X_i \in I_m\}\,.$$

On peut l'écrire formellement de la façon suivante :

$$\mathrm{hist}_n(x) = \sum_{i=1}^{N} C_{n,k} \mathbf{1}_{I_i}(x)\,.$$

En renormalisant, on peut écrire $C_{n,m}/n$ sous la forme d'une moyenne :

$$\frac{C_{n,m}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \in I_m\}} .$$

Et par la loi des grands nombres, lorsque n tend vers l'infini, $\frac{C_{n,m}}{n}$ converge presque sûrement vers l'espérance $\mathbb{E}(\mathbf{1}_{\{X_1 \in I_m\}})$, qui n'est autre que la probabilité $\mathbb{P}(X_1 \in I_m)$:

$$\frac{C_{n,i}}{n} \xrightarrow{p.s.} \mathbb{P}(X_1 \in I_m) .$$

Rappelons par ailleurs, que

$$\mathbb{P}(X_1 \in I_m) = \int_{I_m} p_X(x)dx .$$

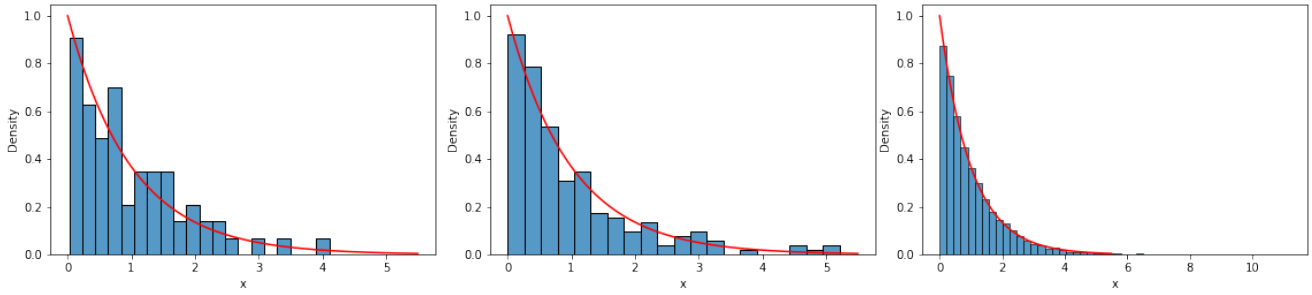
Ainsi, si on a pris soin de choisir M suffisamment grand, de sorte à ce que le segment I_m soit petit, on peut considérer que p_X est à peu près constante sur ce petit intervalle, autrement dit :

$$\mathbb{P}(X_1 \in I_m) \simeq p_X(a + m\frac{b-a}{M})|I_m| ,$$

où $|I_m| = (b-a)/M$ représente la longueur du segment I_m. Ainsi, on obtient que, lorsque n et M sont grands, l'histogramme correctement renormalisé est environ égale à la densité $p_X(x)$:

$$\lim_{M\to\infty}\lim_{n\to\infty} \frac{M}{n(b-a)}\text{hist}_n(x) = p_X(x) .$$

L'histogramme (renormalisé) est donc une façon de visualiser empiriquement la densité de probabilité d'une variables.



(a) (b) (c)

FIGURE 2.3. – Histogramme normalisé associé à un tirage aléatoire de n points selon une loi exponentielle de paramètre 1. On représente sur le même graphique la densité exponentielle de paramètre 1, c'est à dire la fonction e^{-x}. (a) $n = 70, M = 20$, (b) $n = 200, M = 20$, (c) $n = 10000, M = 50$

2.5. Espérance conditionnelle

2.5.1. Cas des variables discrètes

Soient X, Y deux variables aléatoires discrètes (disons à valeurs dans $\mathbb{N}$ pour simplifier). On réalise une expérience aléatoire. On suppose qu'un observateur prend connaissance de la valeur

$X(\omega)$. Il ne connaît pas le résultat ω de l'expérience, et ne peut généralement pas déterminer $Y(\omega)$, en revanche sa connaissance de $X(\omega)$ modifie sa croyance en la valeur prise par Y. *Avant* l'expérience, l'observateur ne peut connaître que la loi de Y, donnée par les coefficients : $\mathbb{P}(Y = \ell)$ pour tout $\ell \in \mathbb{N}$. *Après* l'expérience, si l'observateur s'aperçoit que la réalisation de X vaut k, sa connaissance de Y s'affine en une nouvelle loi, que l'on appelle *loi conditionnelle de Y sachant $X = k$*, et donnée par les coefficients

$$\mathbb{P}(Y = \ell \,|X = k)$$

pour tout $\ell \in \mathbb{N}$, où on rappelle la définition

$$\mathbb{P}(Y = \ell \,|X = k) = \frac{\mathbb{P}(Y = \ell, X = k)}{\mathbb{P}(X = k)},.$$

La loi conditionnelle est bien définie dès lors que $\mathbb{P}(X = k) > 0$.

Exemple 2.29. Soit (X, Y) un couple de v.a. suivant la loi uniforme sur l'ensemble : éléments suivant :

$$T = \{(i, j) \in \mathbb{N} \times \mathbb{N},\ i + 2j \leq 5\}..$$

Cet ensemble comporte 12 éléments, la loi jointe du couple (X, Y) est donc donnée par

$$\mathbb{P}(X = k, Y = \ell) = \frac{1}{12}$$

pour tout couple $(k, \ell) \in T$, et $\mathbb{P}(X = k, Y = \ell) = 0$ pour les autres couples (k, ℓ).

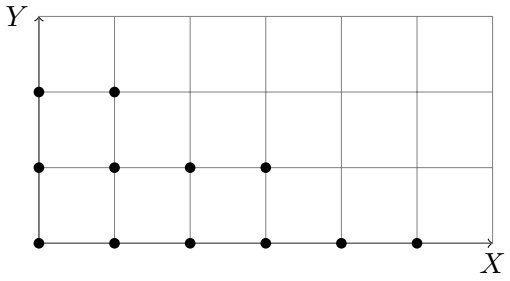


FIGURE 2.4. – L'ensemble T

On note X l'abscisse et Y l'ordonnée. Si l'on sait que $X = 0$, alors Y peut prendre les valeurs 0, 1, 2 tandis que si $X = 2$, Y ne peut prendre que les valeurs 0 et 1. Connaître X donne donc une information sur Y qui se quantifie ainsi :

$$\mathbb{P}(Y = 1 \,|\, X = 0) = \frac{\mathbb{P}(Y = 1,\, X = 0)}{\mathbb{P}(X = 0)} = \frac{1/12}{3/12} = \frac{1}{3}\,.$$

De même,

$$\mathbb{P}(Y = 0 \,|\, X = 0) = \mathbb{P}(Y = 2 \,|\, X = 0) = \frac{1}{3}\cdot$$

On résume ceci en disant que la loi de Y conditionnellement à $X = 0$ est la loi uniforme sur $\{0, 1, 2\}$. Le même raisonnement permet de montrer que la loi conditionnelle de Y sachant $X = 5$ est celle d'une variable aléatoire constante égale à 0.

D'après la formule de Bayes, on peut obtenir la loi conditionnelle de Y sachant X à partir de la loi conditionnelle de X sachant Y :

$$\mathbb{P}(Y = \ell \,|X = k) = \frac{\mathbb{P}(X = k \,|Y = \ell)\,\mathbb{P}(Y = \ell)}{\mathbb{P}(X = k)}\,.$$

Définition 2.30. L'espérance conditionnelle de Y sachant l'événement $\{X = x\}$ est définie par :

$$\mathbb{E}(Y|X = k) = \sum_{\ell \in \mathbb{N}} \ell\, \mathbb{P}(Y = \ell|X = k)\,.$$

Autrement dit, l'espérance conditionnelle est le barycentre de la loi conditionnelle de Y sachant $X = k$. On peut faire trois remarques importantes :

1. Dans le cas où X et Y sont indépendantes, l'observation de l'événement $\{X = k\}$ n'apporte aucune information sur Y, et la loi conditionnelle $\mathbb{P}(Y = \ell|X = k)$ coïncide simplement avec la loi de Y, soit $\mathbb{P}(Y = \ell)$. Cela se traduit sur l'espérance conditionnelle par :

$$\mathbb{E}(Y|X = k) = \mathbb{E}(Y) \text{ si } X, Y \text{ indépendantes.}$$

2. Dans le cas inverse où Y est une fonction de X, disons $Y = h(X)$, alors l'observation de l'événement $\{X = k\}$ détermine entièrement Y comme étant égale à $h(k)$. On a donc,

$$\mathbb{E}(h(X)|X = k) = h(k)\,.$$

3. *L'espérance de l'espérance conditionnelle est l'espérance*, ce qui s'écrit :

$$\mathbb{E}(Y) = \sum_{\ell \in \mathbb{N}} \mathbb{E}(Y|X = k)\mathbb{P}(X = k)\,.$$

Théorème 2.31 (Théorème de transfert)**.** *Pour toute fonction g de X et Y, on a :*

$$\mathbb{E}(g(X, Y)|X = k) = \mathbb{E}(g(k, Y)|X = k) = \sum_{\ell \in \mathbb{N}} g(k, \ell)\,\mathbb{P}(Y = \ell \,|X = k)\,.$$

2.5.2. Cas des variables à densité

Si la loi de X est à densité puisqu'alors l'événement $\{X = x\}$ est de mesure nulle. On ne peut plus calculer de probabilité conditionnelle à l'événement $\{X = x\}$, cela n'a plus de sens. Pourtant, en suivant l'exemple 2.29, on peut très bien imaginer un couple de points répartis au hasard dans le triangle T' donné par la figure 2.5. Dans ce cas, on a fortement envie de dire que, conditionnellement à l'observation que X vaut une certaine valeur x dans l'intervalle $[0, 5]$, Y est distribué selon la loi uniforme sur l'intervalle $[0, 5/2 - x/2]$. Pour formaliser cette idée, on introduit la définition suivante.

Définition 2.32. Soit (X, Y) un couple de v.a. admettant la densité jointe $p_{X,Y}(x, y)$ sur $\mathbb{R}^2$. On appelle *densité conditionnelle* de Y sachant $X = x$, la fonction qui à tout $y \in \mathbb{R}$ associe :

$$p_{Y|X=x}(y) = \frac{p_{X,Y}(x, y)}{p_X(x)}$$

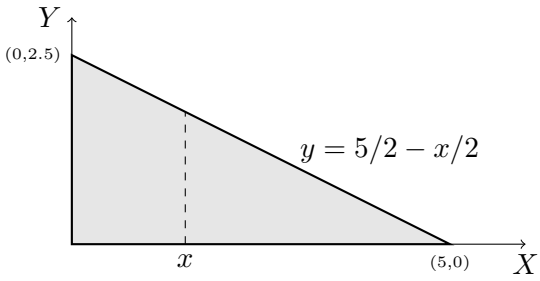


FIGURE 2.5. – L'ensemble T'

où $p_X(x) = \int p_{X,Y}(x,y)dy$ est la densité marginale de X. L'espérance conditionnelle de Y sachant $X = x$ est le barycentre de la densité conditionnelle :

$$\mathbb{E}(Y|X=x) = \int y\, p_{Y|X=x}(y)dy\,.$$

Exemple 2.33. Soit (X,Y) de loi uniforme sur l'ensemble T' défini dans la figure 2.5. La densité jointe est $p_{X,Y}(x,y) = 1/\text{Aire}(T')$ (preuve : $p_{X,Y}$ est par définition constante sur T', et on détermine la constante en écrivant que $1 = \int\int p_{X,Y}(x,y)dxdy$). La densité conditionnelle p_X (définie comme l'intégrale de la densité jointe par rapport à y) est non-nulle uniquement pour $0 < x < 5$. Pour un tel x, on peut obtenir la densité conditionnelle de y sachant $X = x$, en appliquant la définition :

$$p_{Y|X=x}(y) = \frac{1}{\text{Aire}(T')p_X(x)}\mathbf{1}_{T'}(x,y)\,. \tag{2.8}$$

On peut maintenant calculer l'aire de T' et calculer la densité marginale p_X afin d'obtenir l'expression finale. Mais en fait, ce n'est même pas la peine. En effet, il suffit d'observer que, à x fixé, la fonction (2.8) est non-nulle seulement pour $y \in [0, 5/2 - x/2]$ et, sur cet intervalle, est constante par rapport à y. Il s'agit donc de la densité uniforme sur l'intervalle $[0, 5/2 - x/2]$. Autrement dit, on a sans calcul :

$$p_{Y|X=x}(y) = \frac{1}{5/2 - x/2}\mathbf{1}_{[0,5/2-x/2]}(y)\,.$$

On peut alors calculer le barycentre de cette fonction, qui est simplement le point milieu du segment, soit :

$$\mathbb{E}(Y|X=x) = 5/4 - x/4\,.$$

Proposition 2.34. *On a les égalités suivantes, que nous avons déja vues dans le cas discret :*

Si $X \perp\!\!\!\perp Y$, $p_{Y|X=x}(y) = p_Y(y)$ *et* $\mathbb{E}(Y|X=x) = \mathbb{E}(Y)$

$$\mathbb{E}(g(X,Y)\,|X=x) = \int g(x,y)\, p_{Y|X=x}(y)dy \qquad \text{(Théorème de transfert)}$$

$$\mathbb{E}(Y) = \int \mathbb{E}(Y\,|X=x)p_Y(y)dy \qquad \text{(Espérance de l'espérance conditionnelle = espérance)}$$

$$p_{Y|X=x}(y) = \frac{p_{X|Y=y}(x)p_Y(y)}{p_X(x)} \qquad \text{(Formule de Bayes)}$$

2.6. Exercices

Probabilités discrètes

Exercice 2.1. Répondre aux questions suivantes.

1. 25 chevaux participent au tiercé. Combien y a-t-il de tiercés possibles ? (rq : un tiercé est une suite ordonnée)
2. Combien de mots de 5 lettres (pas nécessairement prononçables) peut-on faire avec les 26 lettres de l'alphabet français ?
3. Combien y a-t-il de tirages de 5 cartes dans un jeu de 52 cartes ?
4. Combien de façon a-t-on pour répartir au hasard r boules numérotées de 1 à r, dans n urnes ?

Exercice 2.2. Une classe contient n étudiants. Le professeur parie que deux étudiants au moins ont leur anniversaire le même jour. Que doit valoir n pour que le professeur ait plus d'une chance sur deux de gagner son pari ?

Exercice 2.3. Un jeu télévisé se déroule comme suit. Un candidat est face à trois portes. Une récompense se trouve derrière l'une d'elles. Seul le présentateur connait l'emplacement de la récompense. Le candidat désigne une porte en première intention (sans l'ouvrir pour le moment). Dans un deuxième temps, le présentateur de l'émission ouvre au hasard l'une des deux portes non désignées par le candidat, en prenant garde toutefois de ne jamais ouvrir la porte derrière laquelle est cachée la récompense. On suppose que le candidat désigne la porte 1. On introduit la variable aléatoire $X \in \{1, 2, 3\}$ la position de la récompense et la v.a. Y représentant la porte ouverte par le présentateur.

1. Donner les valeurs de $\mathbb{P}(Y = 2|X = 1)$ et $\mathbb{P}(Y = 3|X = 2)$.
2. Calculer $\mathbb{P}(X = 3|Y = 2)$.
3. Après avoir ouvert la porte Y, le présentateur demande au candidat s'il souhaite main- tenir ou modifier son choix. Que doit faire le candidat pour optimiser ses chances ? Justifier.

Exercice 2.4. Le quart d'une population a été vacciné contre une maladie. Au cours d'une épidémie, on constate qu'il y a parmi les malades un vacciné pour quatre non vaccinés. On sait de plus qu'il y avait un malade sur douze parmi les vaccinés. Quelle était la probabilité de tomber malade pour un individu non vacciné ?

Exercice 2.5. On effectue un suite d'expériences aléatoires, chacune de probabilité de succès p. Cela revient à considérer une suite iid (indépendante et identiquement distribuée) de v.a. de Bernoulli de paramètre $p \in (0, 1)$.

1. On note Y_n le nombre de succès obtenus après n essais. Calculer la loi de Y_n, c'est à dire $\mathbb{P}(Y_n = k)$ pour tout k.
2. Soit T l'instant du premier succès. Exprimer la loi de T.

Exercice 2.6. Une tortue donne naissance à un nombre N de bébés : on note M le nombre de mâles et F le nombre de femelles. On suppose que N est une variable aléatoire suivant une loi

de Poisson de paramètre $\lambda > 0$, c'est à dire :

$$\forall k \in \mathbb{N}, \quad \mathbb{P}[N = k] = \frac{\lambda^k}{k!} e^{-\lambda}$$

et on suppose que chaque bébé a (indépendamment des autres) une chance sur deux d'être une femelle.

1. Calculer la loi du couple (N, F).
   Indication : On pourra écrire $F = Y_1 + Y_2 + \cdots + Y_N$ où $(Y_n)_{n\in\mathbb{N}^*}$ est une suite de variables aléatoires i.i.d., indépendantes de N, et dont on précisera la loi.
2. Calculer la loi de F, puis celle de M.
3. Trouver une relation simple entre M, F et N. Calculer la loi du couple (M, F).
4. Les variables M et F sont-elles indépendantes ?

Variables aléatoires continues

Exercice 2.7. Soit X une v.a. de loi normale $\mathcal{N}(\mu, \sigma^2)$ avec $\sigma^2 > 0$. On rappelle que X admet pour densité :

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Quelle est la densité de $Y = \exp(X)$?

Exercice 2.8. Un ingénieur d'une entreprise de micro-électronique affirme après avoir évalué la robustesse d'un composant A : "63% des composants A ont une durée de vie inférieure à la durée de vie moyenne du composant". Ne devrait-ce pas être par définition 50% ?

Exercice 2.9. Soit A une v.a. exponentielle de paramètre $\alpha > 0$. Déterminer la probabilité pour que le polynôme $x^2 + 2x(A-2) + 4A - 2$ ait toutes ses racines réelles.

Exercice 2.10. (Inégalité de Chebychev-Cantelli) L'inégalité de Chebychev-Cantelli est un raffinement utile de l'inégalité de Bienaymé-Chebychev :

$$\mathbb{P}(X - \mathbb{E}(X) \geq \epsilon) \leq \frac{\mathrm{Var}(X)}{\mathrm{Var}(X) + \epsilon^2}. \tag{2.9}$$

1. On peut supposer sans perte de généralité que $\mathbb{E}(X) = 0$. Justifier que pour tout $\epsilon > 0$,
   $$\epsilon = \mathbb{E}[(\epsilon - X)] \leq \mathbb{E}[(\epsilon - X)\mathbf{1}(X \leq \epsilon)].$$
2. Justifier que
   $$\epsilon^2 \leq \mathbb{E}[(\epsilon - X)^2]\mathbb{P}(X \leq \epsilon) = (\epsilon^2 + \mathrm{Var}(X))\mathbb{P}(X \leq \epsilon),$$
3. Conclure.

Exercice 2.11. Un charcutier fabrique chaque jour x kilos de choucroute : il la vend p euros le kilo, pour un coût de fabrication de $c < p$ euros le kilo. La demande varie d'un jour à l'autre : on la modélise par une variable positive X à densité. On note F_X la *fonction de répartition* de X. S'il lui reste de la choucroute invendue en fin de journée, il l'offre gratuitement à une association. On note P le profit réalisé en une journée par la vente de choucroute (*n.b.* : il peut arriver que $P(\omega) < 0$).

1. Exprimer P en fonction de x, c, p et X.
2. Montrer que
$$\mathbb{E}(P) = \alpha x - \beta \int_0^x F_X(t)dt$$
où α et β sont deux constantes que l'on exprimera en fonction de p et c.
3. Quelle quantité doit produire le charcutier pour maximiser l'espérance de son profit ?
4. Trouver un nombre a tel que le profit est plus grand que a avec probabilité $0,95$. On exprimera a en fonction de c, p et F_X.
5. Application : Exprimer a lorsque X suit une loi exponentielle de paramètre $\lambda > 0$ et lorsque le charcutier produit la quantité optimale x prescrite à la question 2.

Exercice 2.12. On considère un patient atteint d'une certaine pathologie.

1. La durée de vie après diagnostic est modélisée par une v.a. X de loi exponentielle de paramètre λ. Quelle est l'espérance de vie après diagnostic ? Donner l'écart-type et la médiane de survie.
2. Un traitement permet d'allonger significativement la durée de vie et les trois quart des patients répondent positivement au traitement. On modélise la durée de vie comme une variable aléatoire Z définie par
$$Z = YB + X(1 - B)$$
où Y suit une loi exponentielle de paramètre $\lambda/2$ et où B est une variable de Bernoulli dont on précisera le paramètre. Les variables X, Y, B sont supposées indépendantes. Calculer l'espérance de vie.

Exercice 2.13. Soient X, Y deux v.a. iid (indépendantes identiquement distribuées) de loi uniforme sur $[0, 1]$. Calculer la probabilité que $X^2 + Y^2 \leq 1$.

Exercice 2.14. Soit (X, Y) un couple de v.a.r. dont la loi jointe est la loi uniforme sur le disque unité $\{(x, y), x^2 + y^2 \leq 1\}$: la densité est donnée par
$$f(x, y) = \alpha \ \mathbf{1}_{\mathcal{C}}(x, y) \qquad \mathcal{C} := \{(x, y), x^2 + y^2 \leq 1\}.$$

1. Que vaut la constante α ?
2. Déterminer la loi marginale de X. Sans calculs, quelle est la loi de Y ?
3. Les v.a. X et Y sont-elles indépendantes ?

Exercice 2.15. Une cerise est placée sur la circonférence d'un gateau rond que l'on partage en deux au hasard en pratiquant deux découpes suivant des rayons. Si on prend la position de la cerise comme origine des angles, les positions U et V des deux coups de couteau sont des variables uniformément réparties sur $[0, 2\pi]$ et indépendantes.

1. Exprimer la taille T de la part contenant la cerise en fonction de U et de V.
2. Calculer son espérance et la probabilité qu'elle soit plus grosse que l'autre.
3. Quelle doit être la position d'un gourmand qui doit choisir entre la part avec la cerise et la part sans la cerise, avant le découpage ?

Conditionnement

Exercice 2.16. Soient X, Y, T des v.a. discrètes à valeurs réelles. Que peut-on dire des quantités suivantes (sous réserves d'hypothèses de sommabilité adéquates) ?

1. $\mathbb{E}(f(T)|T=t)$
2. $\mathbb{E}(XY|T=t)$ si $X=f(T)$
3. $\mathbb{E}(f(X)|T=t)$ avec X et T indépendantes
4. $\mathbb{E}(S_{10}|S_8=s)$ avec $S_n=\sum_{i=1}^n X_i$ et les X_i iid
5. $\mathbb{E}(S_{31}|X_1=x)$ avec $S_n=\sum_{i=1}^n X_i$ et les X_i iid
6. $\mathbb{E}(\Pi_4|\Pi_2=p)$ avec $\Pi_n=\prod_{i=1}^n X_i$ et les X_i iid.

Exercice 2.17. Soit (X,Y) un couple de loi uniforme sur le disque unité. Donner la loi conditionnelle de Y sachant X.

Exercice 2.18 (Estimateur MMSE (Minimum Mean Square Error)). On cherche à estimer un paramètre Θ non-observé, que l'on modélise comme une v.a.r. On réalise une expérience aléatoire qui conduit au résultat X. Un estimateur de Θ est une variable aléatoire fonction du résultat de l'expérience, notée $\hat{\Theta}(X)$ où $\hat{\Theta}:\mathbb{R}\to\mathbb{R}$ est une certaine fonction à déterminer. On définit l'erreur quadratique moyenne par

$$EQM(\hat{\Theta})=\mathbb{E}((\hat{\Theta}(X)-\Theta)^2)\,.$$

On introduit la fonction $\psi(x)=\mathbb{E}(\Theta|X=x)$.

1. Montrer l'identité $EQM(\hat{\Theta})=\mathbb{E}((\hat{\Theta}(X)-\psi(X))^2)+\mathbb{E}((\psi(X)-\Theta)^2)$
2. Quel est l'estimateur optimal au sens de l'EQM ?

Exercice 2.19 (Classifieur de Bayes). On se donne un couple de variables aléatoires (X,Y) telles que Y est à valeurs dans $\{0,1\}$ et X est, pour simplifier, à valeurs dans $\mathbb{N}$. On se donne une fonction $h:\mathbb{N}\to\{0,1\}$, aussi appelée un *classifieur*, et on voudrait choisir h pour que la probabilité que $h(X)\neq Y$ soit aussi faible que possible. On appelle cette probabilité la probabilité d'erreur, et on la note :

$$R(h)=\mathbb{P}(h(X)\neq Y)\,.$$

1. Justifier que $R(h)=\mathbb{E}(\psi_h(X))$ où $\psi_h(x)=\mathbb{P}(Y\neq h(x)\,|X=x)\,.$
2. On pose $\eta(x)=\mathbb{P}(Y\neq 1\,|X=x)\,.$ Montrer que

$$\psi_h(x)=\eta(x)\mathbf{1}_{h(x)=0}+(1-\eta(x))\mathbf{1}_{h(x)=1}\,.$$

3. En utilisant le fait que $h(x)$ ne prend que les valeurs 0 et 1, montrer que que pour tout x, $\psi_h(x)\geq\psi_{h^*}(x)$, où h est définie par ;

$$h^*(x)=\begin{cases}1 & \text{si } \eta(x)\geq 0.5\\ 0 & \text{sinon.}\end{cases}$$

La fonction h^* est appelée le classifieur de Bayes.

Exercice 2.20. À un arrêt de bus, un jour donné, les bus passent à intervalles réguliers de durée T. Dans les conditions idéales, $T = t_0$. Cependant, le trafic routier fait que T est aléatoire, et on peut modéliser la loi de T par une loi Pareto de paramètres (t_0, α), $T \sim \mathcal{P}(t_0, \alpha)$ avec $\alpha > 0$ fixé, c'est-à-dire

$$\forall t \in \mathbb{R}, \quad \mathbb{P}(T > t) = \begin{cases} \left(\frac{t_0}{t}\right)^{\alpha} & \text{si } t \geq t_0 \\ 1 & \text{sinon.} \end{cases}$$

1. Quelle est la fonction de répartition de T ?
2. Quelle est la densité $p_T(t)$ de T ?

Soit n voyageurs attendant à l'arrêt de bus, avec $n \geq 1$. On note X_i le temps d'attente du voyageur i, pour $i = 1, \ldots, n$. On admet que si $T = t$, alors X_i suit une loi uniforme sur $[0, t]$. Autrement dit la densité conditionnelle de X_i sachant $T = t$, est une loi uniforme sur $[0, t]$.

On note $X = (X_1, \ldots, X_n)$. Les v.a. $X_1, \ldots, X_n$ sont indépendantes conditionnellement à $T = t$, c'est-à-dire que la densité conditionnelle de X sachant $\{T = t\}$ est le produit des densités conditionnelles :

$$p_{X|T=t}(x_1, \ldots, x_N) = p_{X_1|T=t}(x_1) \times \cdots \times p_{X_n|T=t}(x_n)\,.$$

3. Donner la densité de $p_{X_1|T=t}(x)$, la densité de X_1 sachant $T = t$.
4. Donner l'espérance du temps d'attente $\mathbb{E}(X_1)$ du premier passager en fonction de $\alpha > 0$.
   Indication : Commencer par calculer $\mathbb{E}(X_1|T = t)$. Pour certaines valeurs de α, $\mathbb{E}(X_1) = +\infty$.
5. Pour $t \geq t_0$, calculer la densité de conditionnelle $p_{X|T=t}(x_1, \ldots, x_n)$ de X sachant $T = t$.
6. Soit $x = (x_1, \ldots, x_n)$ le vecteur des temps d'attente effectifs des n voyageurs. Montrer que la loi de T sachant $X = x$, notée $p_{T|X=x}(t)$, est une loi de Pareto dont on précisera les paramètres (t_n, α_n) en fonction des observations $(x_1, \ldots, x_n)$. On pourra faire intervenir $m(x) = \max_{i=1,\ldots,n} x_i$.
   Indication : Calculer la densité conditionnelle de T sachant $X = x$.
7. La compagnie de bus s'intéresse à la durée moyenne $\mathbb{E}T$. Grâce aux réseaux sociaux, elle a accès aux temps d'attente de n passagers d'un même bus. Quelle est l'espérance conditionnelle de T, sachant cette information ? Autrement dit, que vaut $\mathbb{E}(T|X = x)$?

Exercice 2.21. Soit $\boldsymbol{X}$ un vecteur aléatoire (colonne) sur $\mathbb{R}^d$, d'espérance nulle. Montrer que $\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}(XX^T)$. En déduire que $\mathrm{Cov}(\boldsymbol{X})$ est une matrice symétrique semi-définie positive.

Exercice 2.22. Soit $\boldsymbol{X}$ un vecteur aléatoire sur $\mathbb{R}^d$. Soit A une matrice $m \times d$. Montrer que :

$$\mathrm{Cov}(A\boldsymbol{X}) = A\,\mathrm{Cov}(\boldsymbol{X})A^T\,.$$

Exercice 2.23. Soient X, Y deux v.a. gaussiennes centrées réduites, indépendantes. On pose :

$$\boldsymbol{W} = \begin{pmatrix} 1 & 0 \\ 1 & -1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}\,.$$

Montrer que $\boldsymbol{W}$ est un vecteur gaussien. Calculer son espérance et sa matrice de covariance.

3. Régression linéaire

Avant de décrire le problème de la régression linéaire, nous introduisons, en guise de préliminaire, une quantité courante en statistique, le coefficient de corrélation de Pearson.

3.1. Coefficient de corrélation de Pearson

3.1.1. Définition

Soit $\{(x_1, y_1), \dots, (x_N, y_N)\}$ un ensemble de N couples sur $\mathbb{R}\times\mathbb{R}$.

Définition 3.1. On appelle *moyenne empirique* des points $\{x_1, \dots, x_N\}$ la quantité :

$$\bar{x} := \frac{1}{N}\sum_{i=1}^{N} x_i \,,$$

et $\bar{y}$ est définie de façon similaire.

Définition 3.2. On appelle *coefficient de corrélation empirique* des couples $\{(x_1, y_1), \dots, (x_N, y_N)\}$, ou *coefficient de corrélation de Pearson*, la quantité :

$$r_{x,y} := \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \,.$$

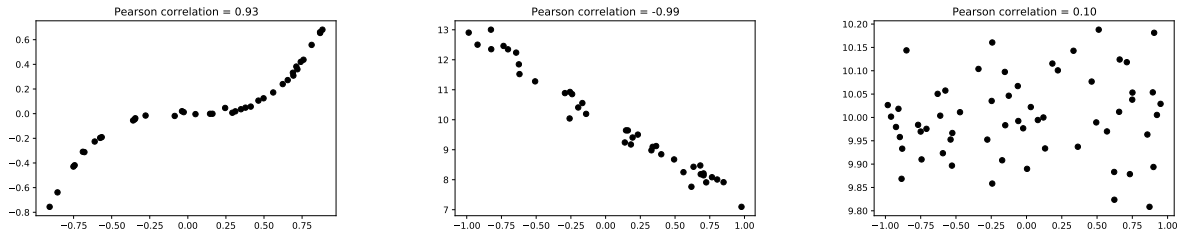
La définition fait sens dès que les x_i ne sont pas tous égaux à $\bar{x}$ (et idem pour y_i), ce que nous supposerons toujours implicitement.

Dans la définition ci-dessus, bien qu'on parle de corrélation, nous n'avons pas défini de variables aléatoires. Il n'y a au départ qu'un simple nuage de points, appelé un N-échantillon. Pourtant, il existe évidemment un lien avec le coefficient de corrélation $\rho_{X,Y}$ défini au premier chapitre dans un cadre probabiliste. Par exemple, si on introduit un vecteur aléatoire (X, Y), suivant la loi uniforme sur l'ensemble $\{(x_1, y_1), \dots, (x_N, y_N)\}$, il apparaît immédiatement que

$$\bar{x} = \mathbb{E}(X) \text{ et } r_{x,y} = \rho_{X,Y} \,.$$

Par conséquent, la proposition 2.25 s'applique. En particulier, le coefficient de Pearson appartient à l'intervalle $[-1, 1]$. Un coefficient de Pearson positif signifie que les valeurs de x_i et y_i ont tendance à être simultanément fortes ou simultanément faibles, par rapport à leur moyenne. Et

dans le cas extrême où $r_{x,y} = 1$, les points x_i, y_i sont même alignés sur une droite de pente positive : $y_i = ax_i + b$ pour certains coefficients $a > 0$ et b. Inversement, si $r_{x,y} = -1$, les points x_i, y_i sont alignés sur une droite, mais dont la pente est cette fois négative. La figure 3.1 fournit quelques exemples.



(a) (b) (c)

FIGURE 3.1. – Coefficient de Pearson pour différents ensembles de points. On a les valeurs suivantes : (a) $r_{x,y} = 0.93$. (b) $r_{x,y} = -0.99$. (c) $r_{x,y} = 0.07$.

3.1.2. Corrélation versus causalité

Exemple 3.3. Deux économistes renommés, C. Reinhart et K. Rogoff, suggèrent dans un article que, dans les pays où la dette publique est supérieure à 90% du produit intérieur brut (PIB), la croissance économique est plus lente. L'article conclut par la recommandation suivante : puisqu'une dette élevée cause un ralentissement de la croissance, il convient de mettre en place des politiques d'austérité, destinées à faire passer le ratio de dette publique sous les 90% du PIB. Quelle erreur fondamentale de raisonnement est commise par les deux économistes ?

De votre environnement professionnel jusqu'à votre sphère privée, vous saurez résister aux interprétations hâtives : corrélation n'implique pas causalité.

Alors comment s'assurer de l'effet causal d'une action sur une quantité observée ? La bonne méthode (à condition qu'on puisse l'appliquer) est l'A/B test. Sans doute connaissez-vous déjà son principe, mais nous l'expliquons avec un exemple.

Exemple 3.4. Une entreprise de e-commerce propose des recommandations personnalisées aux clients qui se rendent sur son site internet. L'équipe datascience de l'entreprise a mis au point une nouvelle IA, dont elle prétend que les recommandations sont mieux adaptées aux besoins des clients. L'entreprise, qui déteste les risques, doit être certaine que ce nouvel algorithme a un impact positif sur le volume de vente, avant de remiser définitivement son ancien algorithme. L'A/B test consiste à diviser les clients en deux groupes A et B : le groupe A est constitué selon un tirage aléatoire uniforme de N individus, dans toute la population de clients. Idem pour le groupe B. La nouvelle IA est testée sur le groupe A, alors que les clients du groupe B utilisent toujours l'ancienne. On compare alors le volume de vente moyen par client dans chacun des deux groupes.

Remarque 3.1. Il n'est pas toujours possible de mettre en œuvre des A/B-tests pour vérifier un

lien causal. Il faut alors recourir à d'autres méthodes statistiques (on parle d'*inférence causale*), qui dépassent le cadre de ce cours.

3.2. Régression linéaire simple

3.2.1. Etude de cas

L'Enquête canadienne sur les mesures de la santé (ECMS) a permis de mesurer la tension artérielle moyenne dans cinq groupes de femmes d'âges différent. Les données sont fournies dans le tableau suivant.

Tension	Âge
102	25
103	35
109	45
119	55
122	65
126	75

TABLE 3.1. – Tension artérielle moyenne de cinq groupes de femmes (source : ECMS)

Nous allons construire un modèle mathématique qui permet de lier la tension d'une femme à son âge. En représentant ces points sur une courbe, il semble que ces points soient approximativement situés sur une droite. Il est donc naturel de chercher une relation du type :

$$\text{tension} \simeq \beta_1 \times \text{âge} + \beta_0$$

où β_0 est l'ordonnée à l'origine, et β_1 la pente de la droite en question. La présence du signe $\simeq$ provient du fait que les points ne sont pas *exactement* situés sur une droite. L'objectif est de trouver la droite (c'est à dire les coefficients β_0, β_1) qui passe "au plus proche" des points (âge, tension). La première étape est donc de nous donner un *critère* qui quantifie à quel point une droite approxime bien ou mal notre nuage de points.

Remarque 3.2. L'âge est ici la *variable explicative* (en anglais : feature, ou input), et la tension est la *variable à expliquer* ou la *réponse* (en anglais : label, ou output).

3.2.2. Critère des moindres carrés

On notera N le nombre de points (x_i, y_i) disponibles, où x_i représente la $i$ème variable explicative et y_i la $i$ème variable à expliquer. Dans notre cas, $N = 6$, et les couples (x_i, y_i) sont donnés par l'ensemble :

$$\{(25, 102), (35, 103), (45, 109), (55, 119), (65, 122), (75, 126)\}.$$

Cet ensemble est le jeu de données (dataset) – on parle aussi de N-échantillon. Chaque point (x_i, y_i) du jeu de données est appelé un échantillon, ou un exemple (en anglais : sample).

Définition 3.5. On appelle *critère des moindres carrés* la fonction

$$J(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 \,.$$

Cette fonction mesure la somme des écarts quadratiques entre la réponse d'un modèle $\beta_1 x_i + \beta_0$ et la variable y_i effectivement observée.

Remarque 3.3. On peut se demander pourquoi avoir élevé l'écart au carré, et pourquoi ne pas choisir, par exemple, la valeur absolue de l'écart. En fait, la valeur absolue aurait été, elle aussi, tout à fait intéressante. Les deux critères pénalisent les erreurs de façon différente (l'erreur quadratique utilisée ici pénalise davantage les grands écarts). Mais au delà de ça, ce qui fait le succès du critère des moindres carrés, c'est surtout la simplicité de sa mise en œuvre numérique, ainsi que l'interprétabilité des résultats.

Proposition 3.6. *Supposons que les x_i ne sont pas tous égaux. On pose $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, et $\bar{y}$ défini de même comme la moyenne des y_i. Alors, le critère des moindre carrés admet pour minimiseur le point $(\hat{\beta}_0, \hat{\beta}_1)$ défini par :*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Le point $(\hat{\beta}_0, \hat{\beta}_1)$ est appelé l'estimateur des moindres carrés.

Démonstration. En cours. □

On remarquera le lien naturel qui existe entre la pente $\hat{\beta}_1$ du modèle obtenu et le coefficient de corrélation de Pearson $r_{x,y}$. Une fois l'estimateur calculé, nous avons notre modèle définitif. Dans le cas de la Table 3.1, il s'agit de :

$$\hat{y}(x) = 0.53x + 86.79 \,,$$

où x représente l'âge et $\hat{y}(x)$ est la tension prédite. La fonction $\hat{y}(\cdot)$ est appelée le *prédicteur*. Dans le contexte du machine learning, la prédiction est une tâche très courante : l'objectif est souvent de prédire une réponse y lorsque l'on observe seulement une certaine variable explicative x. Pour cela, on apprend un modèle à partir d'un *jeu de données labellisé*, c'est à dire d'un N-échantillon constitué de couple (x_i, y_i) pour lesquels la réponse y_i a bien été mesurée. Nous verrons des exemples, plus tard dans ce cours.

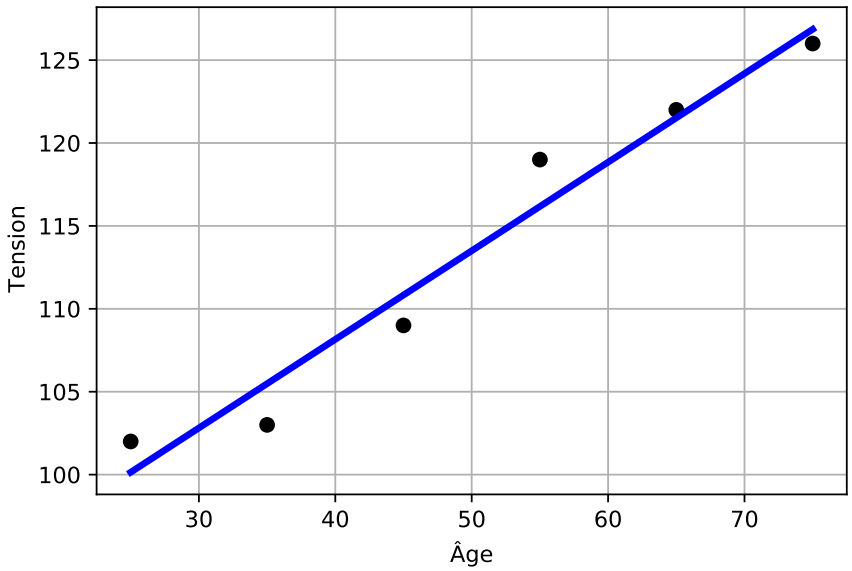


FIGURE 3.2. – Tension en fonction de l'âge chez les femmes. Les points sont les données de l'ECMS. La droite est le prédicteur.

3.3. Régression linéaire multiple

3.3.1. Modèle

On cherche à expliquer une réponse $y \in \mathbb{R}$ en fonction d'un vecteur $\boldsymbol{x} \in \mathbb{R}^d$, de la forme $\boldsymbol{x} = (x_1, \ldots, x_d)^T$. Les variables $x_1, \ldots, x_d$ s'appellent les variables explicatives, ou les régresseurs (*features*, en anglais).

Exemple 3.7. On cherche à prédire la valeur y d'une action, en fonction des variables explicatives suivantes :

- x_1 = le taux d'intérêt de la Banque Centrale Européenne
- x_2 = le coût du baril de pétrole brut
- x_3 = l'indice boursier S&P-500.

On cherche une relation affine du type :

$$y \simeq \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d \,.$$

Pour cela, on dispose d'un N-échantillon :

$$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$$

où chaque couple $(\boldsymbol{x}_i, y_i)$ est un élément de $\mathbb{R}^d \times \mathbb{R}$. Chaque vecteur $\boldsymbol{x}_i$ est caractérisé par ses composantes, que nous notons $x_{i,1}, \ldots, x_{i,d}$, soit :

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,d} \end{pmatrix} .$$

Remarque 3.4. S'il y a d variables explicatives, le modèle comporte donc $d+1$ paramètres. Cela est dû à la présence de l'ordonnée à l'origine β_0.

3.3.2. Critère des moindres carrés

On pose

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} .$$

Le vecteur $\boldsymbol{\beta}$ est de dimension $d+1$. Il contient les paramètres inconnus, que nous devons déterminer à l'aide du N-échantillon. Pour cela, nous minimisons le critère des moindres carrés, défini ci-dessous comme la somme des erreurs quadratiques entre la réponse prédite par le modèle et la variable à expliquer.

Définition 3.8. On appelle critère des moindres carrés la fonction $J : \mathbb{R}^{d+1} \to \mathbb{R}$ définie par :

$$J(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_{i,1} + \cdots + \beta_d x_{i,d})^2 .$$

A partir de maintenant, nous allons utiliser des notations matricielles pour simplifier les écritures. Notamment, on remarque que :

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d = (1, x_{i,1}, \ldots, x_{i,d})\, \boldsymbol{\beta} .$$

Ainsi, on obtient que

$$J(\boldsymbol{\beta}) = \|\boldsymbol{y} - \Phi\boldsymbol{\beta}\|^2$$

où $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ et où Φ est la matrice $N \times (d+1)$ donnée par :

$$\Phi = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & & & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,d} \end{pmatrix} \tag{3.1}$$

Théorème 3.9. *Supposons que* $\operatorname{rang}(\Phi) = d+1$. *Alors J admet un unique minimiseur donné par :*

$$\hat{\beta} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{y} .$$

Le point $\hat{\beta}$ est appelé l'estimateur des moindres carrés.

Dans le cas $d = 1$, on retrouve bien l'estimateur des moindres carrés donné par la proposition 3.6. Etant donné une nouvelle variable $\boldsymbol{x} \in \mathbb{R}^d$, le prédicteur est :

$$\hat{y}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_d x_d .$$

3.3.3. Et ensuite ?

Le travail du statisticien ne s'arrête pas au calcul de l'estimateur des moindres carrés. On doit répondre aux questions suivantes.

1. **Le modèle linéaire est-il bien fondé ?** Il existe au moins trois façons de répondre à cette question. La plus simple est d'utiliser l'oeil : on trace les nuages de points pour vérifier le comportement linéaire de la réponse en fonction des variables explicatives, on trace les résidus pour s'assurer de leur faible amplitude. La seconde est de calculer certaines mesures d'adéquation, comme le *coefficient R^2 de détermination*. Nous n'avons pas le temps de développer ces notions dans le cours. La troisième est d'essayer d'autres modèles, et de comparer les erreurs de prédiction.
2. **Y a-t-il des données aberrantes ?** Ces données correspondent à des points pour lesquels le résidu $e_i = y_i - \hat{y}_i$ est anormalement élevé. On peut tracer les résidus pour détecter les valeurs aberrantes, ou utiliser un test permettant de les détecter automatiquement.
3. **Peut-on fournir des intervalles de confiance ?** Dans l'étude de cas de la table 3.1 (ECMS), nous avons calculé $\hat{\beta}_1 = 0.53$, ce qui indique une croissance de la tension en fonction de l'âge avec une pente de 0.53. Mais quelle confiance a-t-on en ce 0.53 ? Peut-être que si l'enquête avait été menée auprès de plus de sujets, avec plus d'âges différents, nous aurions trouvé un autre résultat, comme $\hat{\beta}_1 = 0.6$? Ce qui va compter pour le statisticien, ce n'est pas la valeur de $\hat{\beta}_1$, c'est de pouvoir affirmer que, avec une forte probabilité, la "vraie" pente se trouve dans un certain intervalle $[a, b]$. Un tel intervalle s'appelle un intervalle de confiance. Son calcul nécessite de placer le problème de l'estimation dans un cadre probabiliste, ce que nous n'avons pas fait jusqu'ici. Nous y reviendrons donc plus tard dans ce polycopié, et la notion d'intervalle de confiance deviendra plus claire à ce moment là.
4. **Quelles sont les variables qui expliquent "réellement" la réponse ?** On a souvent de nombreuses variables explicatives, et toutes ne sont pas forcément utiles. Naïvement, on pourrait dire que si $\hat{\beta}_i$ est proche de zéro, c'est que la $i$ème variable explicative a peu d'influence. Malheureusement, ce raisonnement est trop simpliste car pour répondre proprement, il nous faut aussi la confiance que l'on a dans la valeur du $\hat{\beta}_i$. En fait, formaliser correctement ce problème revient à l'écrire sous la forme d'un test d'hypothèse, permettant de valider ou d'invalider la dépendance en la $i$ème variable. Nous y reviendrons plus tard dans ce cours. Un problème très lié à la question ci-dessus est celui de la *sélection de modèle*. Sur les d variables explicatives de départ, le statisticien cherche souvent à ne conserver qu'un plus petit nombre de variables importantes. Le modèle obtenu est ainsi plus simple, réduit aux variables qui portent réellement l'information utile. En outre, comme vous le verrez en machine learning, les modèles plus simples présentent en général de meilleures propriétés de généralisation, c'est-à-dire que le modèle produira des erreurs de prédiction plus faible quand il sera utilisé sur de nouvelles données.

Dans votre vie professionnelle, lorsque vous aurez à analyser des données, vous serez amenés à vous poser les questions ci-dessus.

3.4. Exercices

Exercice 3.1. Montrer que, dans le cas $d = 1$, l'estimateur des moindres carrés $\hat{\beta} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{y}$ donné par le théorème 3.9 est bien identique à l'estimateur des moindres carrés donné par la

proposition 3.6.

Exercice 3.2. Dans le chapitre sur la régression linéaire multiple, nous avons supposé que la matrice Φ est de rang $d+1$. Qu'est-ce que cela sous-entend sur le rapport entre le nombre de régresseurs et le nombre N d'observations disponibles ?

Exercice 3.3. On considère le modèle de Volterra suivant :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

On considère un N-échantillon (avec $N \geq 3$)

1. Ecrire le modèle sous forme matricielle.
2. Résoudre les moindres carrés. Quelle propriété sur le N-échantillon assure la pseudo-inversion de la matrice. (Indication : regarder les propriétés d'une matrice de Vandermonde).

4. Modèle paramétrique

Dans ce chapitre, nous faisons une avancée majeure : nous allons nous donner un cadre probabiliste pour décrire les données. Dans le chapitre précédent, les données n'étaient que des points. Dans ce chapitre, ce seront des réalisations de certaines variables aléatoires. Cela va ouvrir beaucoup de possibilités, comme, entre autres, le calcul d'intervalles de confiance et le calcul d'erreur d'estimation.

Nous allons commencer par décrire un cadre abstrait, qui fixe le contexte de ce qu'on appelle l'estimation statistique. Nous verrons très vite des exemples qui rendront les choses plus concrètes, mais il faut en passer d'abord par cette abstraction : elle est nécessaire afin de bien comprendre ce que l'on fait, et de ne pas commettre d'erreur profonde de raisonnement. Ce cadre abstrait se nomme le *modèle paramétrique*.

Nous vous conseillons de bien lire le cadre formel. Les exemples qui suivent éclairciront les choses, et vous pourrez relire alors le cadre formel qui vous semblera alors bien moins abstrait.

4.1. Cadre formel

On considère N variables aléatoires réelles $Y_1, \dots, Y_N$, sur un univers Ω, muni d'une probabilité $\mathbb{P}$. L'utilisation d'une lettre majuscule indique bien que ce sont des variables aléatoires (nous utiliserons toujours la convention : majuscule = variable aléatoire, minuscule = variable déterministe). On note :

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$$

le vecteur aléatoire sur $\mathbb{R}^N$.

On considère un observateur : c'est vous, le statisticien. L'observateur a accès à une *réalisation* $\boldsymbol{Y}(\omega)$ pour une certaine issue ω de l'expérience aléatoire. La loi de $\boldsymbol{Y}$ est inconnue de l'observateur. Toutefois, l'observateur fait une *hypothèse* sur cette loi. Il suppose que la loi fait partie d'une certaine famille, indexée par un certain vecteur $\boldsymbol{\theta} \in \mathbb{R}^d$, que l'on appelle le *paramètre*. Cette famille de lois s'appelle le *modèle* : elle relève du choix du statisticien. L'objectif de l'observateur est alors de déterminer la valeur du paramètre $\boldsymbol{\theta}$ qui explique le mieux possible la réalisation $\boldsymbol{Y}(\omega)$. Cette valeur est de la forme :

$$\hat{\boldsymbol{\theta}}(\omega) = \Theta(\boldsymbol{Y}(\omega)) \,.$$

La variable aléatoire $\hat{\boldsymbol{\theta}}$ s'appelle l'estimée. C'est une v.a. sur $\mathbb{R}^d$. La fonction $\Theta : \mathbb{R}^N \to \mathbb{R}^d$ s'appelle l'estimateur. Cette fonction représente la méthode ou l'algorithme utilisé par le statisticien pour produire l'estimée à partir de son observation $\boldsymbol{Y}(\omega)$.

Cas de variables à densité

Le modèle $\mathcal{P}$ choisi par le statisticien consiste en une famille de densités :

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in \mathbb{R}^d\}$$

où pour chaque $\boldsymbol{\theta}$ possible, $p_{\boldsymbol{\theta}}$ est une densité de probabilité sur $\mathbb{R}^N$. La plupart du temps (en tout cas, ce sera toujours le cas dans ce cours), le statisticien fait l'hypothèse que les observations $Y_1, \dots, Y_N$ sont indépendantes. Autrement dit, chaque densité $p_{\boldsymbol{\theta}}$ sera toujours supposée s'écrire comme le produit des densités marginales :

$$\forall (y_1, \dots, y_N) \in \mathbb{R}^N,\ p_{\boldsymbol{\theta}}(y_1, \dots, y_N) = p_{1,\boldsymbol{\theta}}(y_1) \times \cdots \times p_{N,\boldsymbol{\theta}}(y_N)\,,$$

où on a appelé $p_{1,\boldsymbol{\theta}}, \dots, p_{N,\boldsymbol{\theta}}$ les densités marginales (supposées !) de $Y_1, \dots, Y_N$ respectivement. Choisir un modèle $\mathcal{P}$ revient donc à choisir N densités de probabilités $p_{1,\boldsymbol{\theta}}, \dots, p_{N,\boldsymbol{\theta}}$, qui dépendent d'un certain paramètre $\boldsymbol{\theta}$.

Cas de variables discrètes

Si les variables aléatoires $Y_1, \dots, Y_N$ sont à valeurs dans $\mathbb{N}$, le statisticien se donne pour modèle une certaine famille de lois sur $\mathbb{N}^N$:

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in \mathbb{R}^d\}$$

où cette fois $p_{\boldsymbol{\theta}}$ n'est plus une densité de probabilité, mais une loi discrète de la forme :

$$\forall (k_1, \dots, k_N) \in \mathbb{N}^N,\ p_{\boldsymbol{\theta}}(k_1, \dots, k_N) = p_{1,\boldsymbol{\theta}}(k_1) \times \cdots \times p_{N,\boldsymbol{\theta}}(k_N)\,,$$

où $p_{1,\boldsymbol{\theta}}, \dots, p_{N,\boldsymbol{\theta}}$ sont des lois sur $\mathbb{N}$. On a raccourci la notation $p_{\boldsymbol{\theta}}(y_1 = k_1, \dots, y_N = k_N)$ en $p_{\boldsymbol{\theta}}(k_1, \dots, k_N)$ avec les k_n appartenant à l'ensemble discret des valeurs possibles pour bien montrer son aspect discret.

Remarque 4.1. Soyons rigoureux avec les notations ! Ne confondons pas :

- $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$ la variable aléatoire ;
- $\boldsymbol{Y}(\omega) = (Y_1(\omega) \dots, Y_N(\omega))^T$ la réalisation qui est effectivement observée par le statisticien lors d'une certaine expérience aléatoire ;
- $\boldsymbol{y} = (y_1, \dots, y_N)^T$ qui est une variable muette, un point de $\mathbb{R}^N$.

Remarque 4.2. Nous utilisons la même notation $p_{\boldsymbol{\theta}}$ dans le cas à densité et dans le cas discret. Mais il faut bien comprendre que dans le premier cas, les fonctions $p_{1,\boldsymbol{\theta}}(y), \dots, p_{N,\boldsymbol{\theta}}(y)$ sont des densités de probabilité sur $\mathbb{R}$, comme par exemple des densités gaussiennes. Alors que dans le second cas, les fonctions $p_{1,\boldsymbol{\theta}}(k), \dots, p_{N,\boldsymbol{\theta}}(k)$ sont des lois sur $\mathbb{N}$, comme par exemple des lois de Bernoulli ou de Poisson.

Dans certains cas de figures, on peut imposer une contrainte sur le paramètre $\boldsymbol{\theta}$, en stipulant que $\boldsymbol{\theta}$ est seulement autorisé à vivre dans une certaine région $D \subset \mathbb{R}^d$. Par exemple, $D = [0, +\infty)^d$ si on a une contrainte de positivité. En toute généralité, on a donc finalement la définition suivante.

Définition 4.1 (Modèle paramétrique)**.** Un modèle paramétrique est une famille

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in D\}$$

où D est un sous-ensemble de $\mathbb{R}^d$, et où pour tout $\boldsymbol{\theta} \in D$, $p_{\boldsymbol{\theta}}$ est une densité sur $\mathbb{R}^N$ (dans le cas à densité) ou une loi discrète sur $\mathbb{N}^N$ (dans le cas discret). L'entier d est la *dimension* du paramètre $\boldsymbol{\theta}$.

Pour chaque valeur de $\boldsymbol{\theta}$, on introduit une probabilité $\mathbb{P}_{\boldsymbol{\theta}}$ sur l'univers Ω telle que, sous la probabilité $\mathbb{P}_{\boldsymbol{\theta}}$, le vecteur aléatoire $\boldsymbol{Y}$ est de loi $p_{\boldsymbol{\theta}}$. Par exemple, dans le cas à densité, pour tout $H \subset \mathbb{R}^N$,

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Y} \in H) = \int \cdots \int_H p_{\boldsymbol{\theta}}(y_1, \ldots, y_N) dy_1 \ldots dy_N \,.$$

On note de même $\mathbb{E}_{\boldsymbol{\theta}}$ l'espérance associée, c'est à dire que pour une fonction $g : \mathbb{R}^N \to \mathbb{R}$ quelconque :

$$\mathbb{E}_{\boldsymbol{\theta}}(g(\boldsymbol{Y})) = \int \cdots \int g(y_1, \ldots, y_N) p_{\boldsymbol{\theta}}(y_1, \ldots, y_N) dy_1 \ldots dy_N \,.$$

Remarque 4.3. La probabilité $\mathbb{P}$ que nous nous sommes fixés au début, détermine la "vraie" loi des observations $\boldsymbol{Y}$. Cette vraie loi est inconnue (on ne sait rien de $\mathbb{P}$, la seule chose que nous connaissons, c'est le N-échantillon). En se donnant un modèle paramétrique, le statisticien se donne une infinité de mesures possibles, toutes de la forme $\mathbb{P}_{\boldsymbol{\theta}}$, où $\boldsymbol{\theta}$ décrit un ensemble D. Il va chercher la "vraie" loi au sein de la famille de lois qu'il s'est donné. Ainsi, si le statisticien choisit un certain modèle paramétrique, c'est parce qu'il a des raisons de penser que la "vraie" loi des observations $\boldsymbol{Y}$ appartient à la famille $\mathcal{P}$, ou en tout cas en est proche. Ces raisons proviennent de son inspection préalable des données, de son travail de visualisation des données. Si le statisticien a choisi un modèle paramétrique raisonnable, on peut alors supposer qu'il existe une certaine valeur $\boldsymbol{\theta}^*$, inconnue, telle que $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*}$, c'est à dire que $\boldsymbol{Y}$ suit la loi $p_{\boldsymbol{\theta}^*}$.

Dans la suite, nous allons étudier deux cas de modèles très répandus.

4.2. Modèle de Bernoulli

4.2.1. Etude de cas

Afin d'estimer la part de fumeurs dans la population française, le ministère de la santé effectue un sondage auprès d'un échantillon de $N = 500$ personnes (voir la table 4.1).

Indice	Réponse
1	0
2	0
3	0
4	1
5	0
⋮	⋮
499	0
500	1

TABLE 4.1. – Résultat d'un sondage. L'indice représente la personne interrogée, la valeur 0 indique la réponse "non" à la question "Etes-vous fumeur/fumeuse ?", la valeur 1 indique la réponse "oui".

4.2.2. Modèle paramétrique

Le statisticien formalise le problème de la manière suivante. La population française est divisée en deux classes, la classe "fumeur" des individus qui se disent fumeurs, et la classe "non-fumeur". Appelons θ^* le rapport entre le nombre de fumeurs divisé par la taille N_{pop} de la population. L'objectif est de déterminer θ^*.

Un sondage consiste à choisir N individus au sein de la population, et à mesurer leur réponse $0/1$ à la question posée. Le statisticien fait l'hypothèse que l'institut de sondage a choisi ces individus de manière aléatoire, indépendante et uniforme au sein de la population.

Remarque 4.4. L'hypothèse d'un échantillonnage iid uniforme pourrait être discutée. Imaginons par exemple que l'institut de sondage parisien ait, par facilité, interrogé des personnes de la région parisienne ? Quel problème cela peut-il poser ?

Le statisticien modélise les réponses des sondés comme étant une réalisation de N variables aléatoires $Y_1, \dots, Y_N$. Si l'hypothèse d'un échantillonnage iid uniforme est satisfaite, la probabilité que la réponse Y_i du $i$ème individu soit 1 ("oui") est égale à la probabilité que cet individu ait été choisi au sein de la classe "fumeur", soit θ^*.

Le modèle paramétrique naturel est donc une famille de lois de Bernoulli de paramètre $\theta \in [0,1]$. Autrement dit :

$$p_{1,\theta}(k) = \cdots = p_{N,\theta}(k) = \begin{cases} \theta & \text{si } k = 1 \\ 1-\theta & \text{si } k = 0\,. \end{cases} \tag{4.1}$$

Le modèle paramétrique est donc :

$$\mathcal{P} = \{k \mapsto \theta^k (1-\theta)^{1-k} \,:\, \theta \in [0,1]\}\,.$$

Le paramètre θ est ici un scalaire, c'est pourquoi nous ne l'écrivons pas en gras.

4.2.3. Estimateur de la moyenne empirique

Etant donné l'observation des variables $Y_1, \ldots, Y_N$ donnée par la table 4.1, le statisticien définit naturellement l'estimée du paramètre θ par :

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} Y_i \,.$$

Afin d'alléger les notations, nous ne mettons pas la dépendance à l'issue ω lorsqu'il n'y pas d'équivoque. Le paramètre θ représente la probabilité qu'une personne réponde positivement au sondage, l'estimée $\hat{\theta}$ est le nombre moyen de personnes ayant répondu positivement, calculée sur le N-échantillon. Autrement dit, l'estimée est la fréquence de réponses positives au sondage.

Pourquoi cet estimateur plutôt qu'un autre ? Certes, nous pourrions justifier théoriquement que la moyenne empirique est un bon estimateur, et même qu'il est optimal, au sens d'un certain critère d'erreur quadratique que nous définirons plus bas. Mais à notre stade, une telle justification est superflue. Un tel estimateur provient tout simplement du bon sens. Après tout, que pourrait on raisonnablement faire d'autre ?

Biais

Le biais de l'estimateur est défini, pour chaque valeur de θ, par :

$$b(\theta) = \mathbb{E}_\theta(\hat{\theta}) - \theta \,.$$

L'espérance $\mathbb{E}_\theta(\hat{\theta})$ est l'espérance de l'estimée $\hat{\theta}$, sous l'hypothèse que les observations sont distribuées selon la loi $p_{\boldsymbol{\theta}}$. Dans le cas présent :

$$\begin{aligned} \mathbb{E}_\theta(\hat{\theta}) &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_\theta(Y_i) \\ &= \mathbb{E}_\theta(Y_1) \\ &= \theta \,, \end{aligned}$$

car, sous $\mathbb{P}_\theta$, les v.a. Y_i sont identiquement distribuées selon la loi de Bernoulli de paramètre θ. On remarque donc que :

$$b(\theta) = 0$$

quelque soit la valeur de θ. On dit que l'estimateur est *sans biais*. Cela veut juste dire que si on traçait la densité de probabilité de $\hat{\theta}$ (ce qui n'est pas possible car il faudrait une infinité d'observations ; en pratique on n'a accès qu'à une estimée de cette densité via un histogramme), elle serait centrée autour de la vraie valeur recherchée.

Variance et erreur quadratique moyenne

La variance de l'estimateur sous $\mathbb{P}_\theta$ est donnée par :

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta((\hat{\theta} - \mathbb{E}_\theta(\hat{\theta}))^2) \,.$$

Sous $\mathbb{P}_\theta$, la v.a. $\sum_{i=1}^N Y_i$ suit une loi binomiale de paramètres (θ, N). La variance est donnée par $N\theta(1-\theta)$. Ainsi la variance de l'estimateur est $\mathrm{Var}_\theta(\hat{\theta}) = \theta(1-\theta)/N$.

L'*erreur quadratique moyenne* (EQM) de l'estimateur est définie par

$$EQM_\theta = \mathbb{E}_\theta((\hat{\theta} - \theta)^2) = \mathrm{Var}_\theta(\hat{\theta}) + b(\theta)^2 \,.$$

Naturellement, lorsque l'estimateur est non-biaisé (comme c'est le cas ici), l'erreur quadratique moyenne coïncide avec la variance de l'estimateur. On a donc :

$$EQM_\theta = \frac{\theta(1-\theta)}{N} \,.$$

L'EQM est un critère de performance de l'estimateur. Surtout, l'EQM permet de comparer deux estimateurs entre eux : on préférera utiliser l'estimateur dont l'EQM est plus petite pour tout θ.

4.2.4. Intervalle de confiance

A partir du sondage, on ne peut évidemment pas répondre à la question "Quelle est la part θ^* de fumeurs dans la population ?". On peut toutefois fournir une réponse statistique :

Grâce au sondage, on peut affirmer qu'avec une probabilité de 0.95, que la part de θ^* de fumeurs est comprise entre 30.5 et 31.5 pourcents.

Autrement dit, on répond à la question en donnant non pas une estimée $\hat{\theta}$ de θ^*, mais un *intervalle* qui contient θ^*, avec forte probabilité, choisie arbitrairement comme étant 0.95 (les valeurs les plus courantes sont 0.95 et 0.99). Cet intervalle $[30.5, 31.5]$ s'appelle un *intervalle de confiance de niveau 95%*.

Une astuce pour construire un intervalle de confiance à 95%, est de considérer l'estimateur $\hat{\theta}$. En utilisant l'inégalité de Bienaymé-Chebychev (2.6), on a pour tout $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_\theta(|\hat{\theta} - \theta| > \epsilon) &\leq \frac{\mathrm{Var}_\theta(\hat{\theta})}{\epsilon^2} \\ &= \frac{\theta(1-\theta)}{N\epsilon^2} \\ &\leq \frac{1}{4N\epsilon^2}\end{aligned}$$

où on a utilisé que $\theta(1-\theta) \leq 1/4$ pour tout $\theta \in [0,1]$. On choisit ϵ de telle sorte que $\frac{1}{4N\epsilon^2} = 0.05$, soit $\epsilon = \frac{1}{\sqrt{0.2N}}$. On obtient :

$$\mathbb{P}_\theta(|\hat{\theta} - \theta| \leq \frac{1}{\sqrt{0.2N}}) \geq 0.95 \,.$$

En particulier, pour $\theta = \theta^*$ (la vraie part de fumeurs dans la population), on obtient :

$$\mathbb{P}_{\theta^*}\left(\theta^* \in \left[\hat{\theta} - \frac{1}{\sqrt{0.2N}}, \hat{\theta} + \frac{1}{\sqrt{0.2N}}\right]\right) \geq 0.95 \,. \tag{4.2}$$

L'intervalle $\left[\hat{\theta} - \frac{1}{\sqrt{0.2N}}, \hat{\theta} + \frac{1}{\sqrt{0.2N}}\right]$ est un intervalle aléatoire. Ses bornes dépendent des observations $Y_1, \ldots, Y_N$. On l'appelle intervalle de confiance de niveau 0.95%. Autrement, sous l'hypothèse que la vraie loi des données est bien p_{θ^*} (c'est à dire sous l'hypothèse que le choix des sondés est iid uniforme dans la population), on peut affirmer que θ^* est dans l'intervalle en question, avec une probabilité de se tromper inférieure à 0.05.

Quand le statisticien fournit un intervalle de confiance, il cherche naturellement l'intervalle le plus court possible (on dit "le plus exact"). Or l'inégalité de Bienaymé-Chebychev que nous avons utilisée pour calculer l'intervalle de confiance est assez grossière. Les exercices 4.1 et 4.2 montrent que l'on peut calculer des intervalles de confiance plus exacts, en utilisant des inégalités plus fines.

4.3. Modèle linéaire gaussien

4.3.1. Etude de cas

Nous utilisons ici des résultats d'une expérience menée par Moore (1975) et analysée par Chatterjee et Hadi (1986). Ces données ont été collectées dans un bio-réacteur, pendant une période de 220 jours. Les données sont reproduites dans le tableau 4.2. Les variables mesurées sont : Y = log(demande d'oxygène) (g/min) ; x_1 = demande d'oxygène biologique (g/litre). x_2 = quantité totale d'azote, g/litre ; x_3 = quantité totale de matière solide, g/litre, x_4 = quantité totale de solides volatils, g/litre ; et x_5 =, demande chimique d'oxygène, g/litre. L'objectif est comme d'habitude d'expliquer la réponse en fonction des *régresseurs* $x_1, \ldots, x_5$. Pour cela, nous allons effectuer une régression linéaire multiple, comme nous avons appris à le faire au chapitre précédent.

Mais, afin d'aller plus loin dans l'interprétation des résultats, nous allons faire une hypothèse probabiliste sur les données. Autrement dit, nous allons commencer par nous fixer un modèle paramétrique $\mathcal{P}$.

Afin de nous placer dans le cadre formel défini au paragraphe 4.1, nous allons donc supposer que le vecteur des réponses observées est une *réalisation* d'un vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$, soit :

$$\boldsymbol{Y}(\omega) = \begin{pmatrix} Y_1(\omega) \\ Y_2(\omega) \\ \vdots \\ Y_{20}(\omega) \end{pmatrix} = \begin{pmatrix} 0.0016 \\ 0.0009 \\ \vdots \\ -0.0000 \end{pmatrix} .$$

Mais comme on le voit dans la table 4.2, l'observation ne se limite pas aux réponses. On observe également, pour chacune des vingt mesures, les valeurs de 5 régresseurs. Nous posons :

$$\begin{aligned} \boldsymbol{x}_1^T &= (1.1250, 0.2320, 7.1600, 0.0859, 8.9050) \\ &\vdots \\ \boldsymbol{x}_{20}^T &= (0.0790, 0.3340, 2.7770, 0.0719, 2.5990) . \end{aligned}$$

x_1	x_2	x_3	x_4	x_5	Y
1.1250	0.2320	7.1600	0.0859	8.9050	0.0016
0.9200	0.2680	8.8040	0.0865	7.3880	0.0009
0.8350	0.2710	8.1080	0.0852	5.3480	0.0007
1.0000	0.2370	6.3700	0.0838	8.0560	0.0007
1.1500	0.1920	6.4410	0.0821	6.9600	0.0003
0.9900	0.2020	5.1540	0.0792	5.6900	0.0004
0.8400	0.1840	5.8960	0.0812	6.9320	0.0001
0.6500	0.2000	5.3360	0.0806	5.4000	0.0001
0.6400	0.1800	5.0410	0.0784	3.1770	-0.0002
0.5830	0.1650	5.0120	0.0793	4.4610	-0.0002
0.5700	0.1510	4.8250	0.0787	3.9010	0
0.5700	0.1710	4.3910	0.0780	5.0020	0
0.5100	0.2430	4.3200	0.0723	4.6650	-0.0001
0.5550	0.1470	3.7090	0.0749	4.6420	-0.0002
0.4600	0.2860	3.9690	0.0744	4.8400	-0.0004
0.2750	0.1980	3.5580	0.0725	4.4790	-0.0002
0.5100	0.1960	4.3610	0.0577	4.2000	-0.0002
0.1650	0.2100	3.3010	0.0718	3.4100	-0.0004
0.2440	0.3270	2.9640	0.0725	3.3600	-0.0005
0.0790	0.3340	2.7770	0.0719	2.5990	-0.0000

TABLE 4.2. – Données de Moore

Le jeu de données de la table 4.2 est donc compris par le statisticien comme étant de la forme :

$$\{(\boldsymbol{x}_1, Y_1(\omega)), \ldots, (\boldsymbol{x}_N, Y_N(\omega))\},$$

où ω est l'issue de l'expérience aléatoire, où pour tout i, $\boldsymbol{x}_i$ est un vecteur déterministe de $\mathbb{R}^5$, et Y_i est un variable aléatoire sur $\mathbb{R}$, et où $N = 20$.

Remarque 4.5. Insistons sur les notations. La notation $\boldsymbol{x}_i$ est en gras, donc il s'agit d'un vecteur, et en minuscule, donc il s'agit d'une quantité déterministe, qui ne dépend pas de ω. La notation Y_i est en majuscules, c'est donc une variable aléatoire, et n'est pas en gras, donc les valeurs sont scalaires.

Remarque 4.6. Ainsi, la différence majeure de ce chapitre par rapport au chapitre 3 est que les réponses sont supposées être des réalisations de variables aléatoires. Par comparaison au chapitre 3, ce choix de modèle apporte de nouvelles perspectives en termes de construction d'estimateurs et d'interprétation des résultats.

Il est temps de fixer notre choix de modèle paramétrique $\mathcal{P}$.

4.3.2. Modèle homoscédastique

Considérons un ensemble de couples $(\boldsymbol{x}_i, Y_i)$ pour $i = 1, \ldots, N$, où $\boldsymbol{x}_i \in \mathbb{R}^d$ et où Y_i est un v.a.r. Les entrées de chaque vecteur $\boldsymbol{x}_i$ seront notées, comme au chapitre précédent, $\boldsymbol{x}_i =$

$(x_{i,1}, \dots, x_{i,d})^T$. On suppose dorénavant que $N > d+1$.

Le modèle homoscédastique consiste à faire l'hypothèse que les réponses Y_i observées correspondent à une fonction affine du vecteur $\boldsymbol{x}_i$, à laquelle s'ajoute une perturbation gaussienne. Autrement dit :

$$\forall i,\, Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_d x_{i,d} + \varepsilon_i \,, \tag{4.3}$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ est une v.a. gaussienne centrée d'espérance nulle et de variance σ^2. On suppose en outre que les variables $\varepsilon_1, \dots, \varepsilon_N$ sont indépendantes. Dans ce modèle, les v.a. $Y_1, \dots, Y_N$ sont donc indépendantes, mais pas identiquement distribuées. Le vecteur $\boldsymbol{Y} = (Y_1, \dots, Y_N)^T$ s'écrit ainsi :

$$\boldsymbol{Y} = \Phi\boldsymbol{\beta} + \boldsymbol{\varepsilon} \,,$$

où $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T$, où $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^T$ et où Φ est la matrice définie par (3.1), et dont nous rappelons l'expression :

$$\Phi = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,d} \\ \vdots & & & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,d} \end{pmatrix} .$$

Par conséquent, $\boldsymbol{Y}$ est un vecteur gaussien d'espérance $\Phi\boldsymbol{\beta}$ et de matrice de covariance $\sigma^2 I_N$, où I_N est l'identité de taille N :

$$\boldsymbol{Y} \sim \mathcal{N}(\Phi\boldsymbol{\beta}, \sigma^2 I_N) \,. \tag{4.4}$$

Dans cette modélisation, la loi de $\boldsymbol{Y}$ dépend de $d+2$ paramètres : $\beta_0, \dots, \beta_d$ et la variance σ^2. Le vecteur final constituant l'ensemble des paramètres scalaires inconnus est donc :

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \\ \sigma^2 \end{pmatrix} .$$

La densité de probabilité de $\boldsymbol{Y}$ est donnée, pour tout $\boldsymbol{y} \in \mathbb{R}^N$, par :

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) = \frac{1}{(\sqrt{2\pi\sigma^2})^N} \exp\left(-\tfrac{1}{2\sigma^2}\|\boldsymbol{y} - \Phi\boldsymbol{\beta}\|^2\right) . \tag{4.5}$$

Le paramètre $\boldsymbol{\theta}$ décrit l'ensemble $D = \mathbb{R}^{d+1} \times (0, +\infty)$. Le modèle paramétrique est donc finalement donné par :

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} \,:\, \boldsymbol{\theta} \in D\} \,.$$

Remarque 4.7. La variable ε_i est une perturbation stochastique (un peu abusivement, on parle parfois de *bruit*) qui caractérise le fait que les réponses ne s'écrivent pas exactement comme une fonction affine des régresseurs. Le terme homoscédastique traduit le fait que les ϵ_i sont supposés avoir tous la même variance, contrairement à un modèle dit hétéroscédastique, où les variances pourraient dépendre de i.

4.3.3. Estimateur

Le vecteur de paramètres à estimer se décompose en $\boldsymbol{\beta}$ et σ^2. Le paramètre d'intérêt consiste surtout en les coefficients $\beta_1, \dots, \beta_d$ qui vont révéler l'influence des différents régresseurs sur la réponse. Nous choisissons naturellement l'estimateur des moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\Phi^T \Phi)^{-1} \Phi^T \boldsymbol{Y} .$$

Proposition 4.2. *Pour tout $\boldsymbol{\beta}, \sigma^2$, dans le cadre du modèle homoscédastique, on a*

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \frac{\sigma^2}{N}(\Phi^T \Phi)^{-1}).$$

En particulier, l'estimateur des moindres carrés est non biaisé, c'est à dire que $\mathbb{E}_{\boldsymbol{\beta},\sigma^2}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$.

Démonstration. Il suffit de remarquer que $\boldsymbol{Y}$ est un vecteur gaussien d'après l'équation (4.4). L'estimateur $\hat{\boldsymbol{\beta}}$ est une fonction linéaire d'un vecteur gaussien, c'est donc un vecteur gaussien. Il suffit de calculer son espérance et sa covariance pour obtenir le résultat. □

Remarque 4.8. Grâce au théorème de Cramér-Rao, on peut établir que, dans le cas du modèle homoscédastique, l'estimateur des moindres carrés est, parmi tous les estimateurs non-biaisés, celui qui possède la plus faible variance.

4.3.4. Intervalle de confiance

Soit $k = 0, \dots, d$ fixé. Nous voulons fournir un intervalle de confiance à 95% sur le paramètre β_k. Pour rappel, il s'agit d'un intervalle $I(\boldsymbol{Y})$ dont les extrémités sont des variables aléatoires, qui dépendent des observations, et qui sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$, contient β_k avec une probabilité au moins égale à 0.95.

Première approche

Nous allons commencer par une approche simple qui permet de comprendre le mécanisme de la construction d'un intervalle de confiance. Nous affinerons cette approche dans un second temps.

La technique pour déterminer un tel intervalle consiste à inspecter la loi de l'estimateur $\hat{\beta}_k$. Plaçons-nous sous la loi $\mathbb{P}_{\beta,\sigma^2}$ qui est gaussienne en raison du modèle homoscédastique. On peut démontrer le résultat suivant :

$$\hat{\beta}_k \sim \mathcal{N}(\beta_k, \frac{\sigma^2 s_k}{N})$$

où s_k est défini comme le $k$ème coefficient de la diagonale de $(\Phi^T \Phi)^{-1}$. Ce résultat est une conséquence d'un théorème, appelé Théorème de Gauss-Markov, qui n'est pas très compliqué à démontrer, mais que nous omettons par souci d'efficacité (les élèves curieux pourront faire

une recherche sur internet ou interroger leur professeur). On peut se ramener à une loi normale centrée réduite par :

$$\sqrt{\frac{N}{\sigma^2 s_k}}(\beta_k - \hat{\beta}_k) \sim \mathcal{N}(0,1)\,. \tag{4.6}$$

Définissons par $F(x)$ la fonction de répartition de la loi $\mathcal{N}(0,1)$, soit :

$$F(x) := \int_{-\infty}^{x} \tfrac{1}{\sqrt{2\pi}} e^{-t^2/2} dt\,.$$

D'après l'équation (4.6), on a pour tout $a > 0$,

$$\begin{aligned} F(a) - F(-a) &= \mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(-a \le \sqrt{\frac{N}{\sigma^2 s_k}}(\beta_k - \hat{\beta}_k) \le a\right) \\ &= \mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\hat{\beta}_k - a\sqrt{\frac{\sigma^2 s_k}{N}} \le \beta_k \le \hat{\beta}_k + a\sqrt{\frac{\sigma^2 s_k}{N}}\right) \end{aligned}$$

Choisissons a pour que $F(a) - F(-a) = 0.95$. Si Z est une gaussienne centrée réduite, rappelons que

$$F(a) = \mathbb{P}(Z \le a) = \mathbb{P}(-Z \le a) = \mathbb{P}(Z \ge -a) = 1 - F(-a)\,.$$

Par conséquence, on doit choisir a pour que $F(a) = 0.975$. Autrement dit, a est le quantile de niveau 0.975 de la gaussienne centrée réduite. Tout statisticien aguerri sait que ce quantile vaut $a = 1.96$. Nous avons donc démontré que :

$$\mathbb{P}_{\boldsymbol{\beta},\sigma^2}\left(\beta_k \in \hat{\beta}_k \pm 1.96\sqrt{\frac{\sigma^2 s_k}{N}}\right) = 0.95\,.$$

L'intervalle $\hat{\beta}_k \pm 1.96\sqrt{\frac{\sigma^2 s_k}{N}}$ contient donc le paramètre inconnu β_k avec probabilité 0.95. Cet intervalle dépend des données au travers de l'estimée $\hat{\beta}_k$ et au travers du coefficient s_k.

Malheureusement, cet intervalle ne peut pas être considéré comme un intervalle de confiance, car il dépend du paramètre inconnu σ^2. Tout espoir n'est pas perdu : il nous suffit d'estimer σ^2.

Deuxième approche

On introduit l'estimateur :

$$\hat{\sigma}^2 = \frac{\|Y - \Phi\hat{\boldsymbol{\beta}}\|^2}{N - d - 1}\,.$$

On se souvient que le vecteur $\Phi\hat{\boldsymbol{\beta}}$ est le vecteur des prédictions, et donc la différence $Y - \Phi\hat{\boldsymbol{\beta}}$ est le vecteur des résidus. La norme au carré $\|\boldsymbol{Y} - \Phi\hat{\boldsymbol{\beta}}\|^2$ est donc égale au SS_R. Enfin, la division par $N - d - 1$ plutôt que par N assure que $\hat{\sigma}^2$ est un estimateur non biaisé, comme le montre le résultat suivant, c'est à dire que $\mathbb{E}_{\boldsymbol{\beta},\sigma^2}[\hat{\sigma}^2] = \sigma^2$ pour tout $\boldsymbol{\beta}, \sigma^2$.

On rappelle les définitions de la table 2.3. La loi du chi-deux à k degrés de libertés, notée $\chi^2(k)$, est la loi de la somme des carrés de k variables iid gaussiennes centrées réduites. La loi de Student à k degrés de libertés, notée $\mathcal{T}(k)$ est la loi du rapport $\frac{Z}{\sqrt{U/k}}$ où Z est une gaussienne centrée réduite, U suit un chi-deux à k degrés de libertés, indépendante de Z. Le résultat suivant est admis.

Lemme 4.3 (Lemme de Cochran). *Sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$ du modèle homoscédastique, la v.a. $\hat{\sigma}^2$ est indépendante de $\hat{\boldsymbol{\beta}}$, et sa loi est caractérisée par :*

$$(N-d-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N-d-1)$$

En outre, pour tout $k = 0, \dots, d$,

$$\sqrt{\frac{N}{s_k \hat{\sigma}^2}}(\hat{\beta}_k - \beta_k) \sim \mathcal{T}(N-d-1)\,.$$

Ainsi, on peut reprendre le même raisonnement qu'au début du paragraphe, d'une part en remplaçant la variance σ^2 inconnue par son estimée $\hat{\sigma}^2$, et d'autre part en remplaçant la loi $\mathcal{N}(0,1)$ par la loi $\mathcal{T}(N-d-1)$. On a donc démontré le résultat suivant.

Théorème 4.4 (Intervalle de confiance). *Soit $k = 0, \dots, d$. Soit $\alpha \in (0,1)$. Soit q_α le quantile de niveau $(1-\alpha/2)$ de la loi de Student $\mathcal{T}(N-d-1)$. Alors, sous $\mathbb{P}_{\boldsymbol{\beta},\sigma^2}$ du modèle homoscédastique, l'intervalle*

$$\hat{\beta}_k \pm q_\alpha \sqrt{\frac{\hat{\sigma}^2 s_k}{N}}$$

est un intervalle de confiance sur β_k de niveau $1-\alpha$.

4.3.5. Interprétation du modèle

Une fois les intervalles de confiances calculés, la question que l'on se pose est la suivante :

Le kème régresseur x_k contribue-t-il à expliquer la réponse Y ?

Cette question est essentielle pour l'*interprétation* du modèle. En supposant que les données suivent le modèle homoscédastique (4.3) pour certains paramètres inconnus $\boldsymbol{\beta}, \sigma^2$, la question une fois formalisée devient :

Est-ce que le coefficient β_k est non-nul ?

La réponse du statisticien ne sera pas "oui" ou "non" : il répondra en terme de niveau de confiance.

Exemple 4.5. On suppose que l'estimateur des moindres carrés appliqué à un certain modèle conduit aux résultats de la figure 4.3.

On observe que $\hat{\beta}_1 = 0.01$ est petit, et que $\hat{\beta}_2 = 32.5$ est plus grand. Ces seules valeurs ne permettent pas de conclure : elles sont peut-être très incertaines ; elles peuvent aussi être dues à la nature des régresseurs x_1 et x_2 qui peuvent être d'amplitudes très différentes, si les régresseurs n'ont pas été préalablement normalisés. Une meilleure approche consiste à inspecter les intervalles de confiance. Avec probabilité 95%, on sait que $\beta_1 \in [0.099, 0.011]$, or cet intervalle ne contient

k	$\hat{\beta}_k$	Intervalle de confiance à 95%
1	0.01	[0.099, 0.011]
2	32.5	[-127, 188]

TABLE 4.3. – Estimées moindres-carrés et intervalles de confiances (exemple hypothétique).

pas zéro. On peut donc conclure avec au moins 95% de chance d'avoir raison, que β_1 n'est pas nul. Il n'en va pas de même pour le second régresseur : l'intervalle de confiance contient l'origine, donc on ne peut pas conclure à la significativité du second régresseur dans l'explication de la réponse.

p-valeur

Sous l'hypothèse que $\beta_k = 0$, le lemme 4.3 implique que $\sqrt{\frac{N}{s_k \hat{\sigma}^2}}\hat{\beta}_k$ suit une loi de Student à $N - d - 1$ degrés de liberté. Le carré d'une v.a. de Student suit ce que l'on appelle une loi de Fisher $\mathcal{F}(1, N - d - 1)$ à $N - d - 1$ degrés de liberté. La densité de Fisher est donnée dans la table 2.3, mais nous nous intéressons surtout à sa fonction de répartition complémentaire que l'on note

$$\bar{F}(x) = \mathbb{P}(F > x) \quad \text{où } F \sim \mathcal{F}(1, N - d - 1)\,.$$

L'expression exacte de $\bar{F}$ est un peu alambiquée, inutile donc de l'écrire, il faut simplement retenir que cette fonction $\bar{F}$ existe, et qu'elle est disponible dans tout bon logiciel ou librairie de statistique.

Ainsi, si nous posons :

$$F_k := \frac{N}{s_k \hat{\sigma}^2}\hat{\beta}_k^2\,,$$

nous pouvons affirmer que, sous l'hypothèse que $\beta_k = 0$, F_k suit la loi $\mathcal{F}(1, N - d - 1)$. Pour savoir si l'hypothèse $\beta_k = 0$ est plausible ou si elle ne l'est pas, il suffit de comparer la valeur de F_k effectivement calculée à la distribution de Fisher. Est-il vraisemblable que ce F_k soit une réalisation de la loi $\mathcal{F}(1, N - d - 1)$? La quantité

$$p_k := \bar{F}\left(F_k\right)$$

est appelée la *p-valeur associée à l'hypothèse* $\beta_k = 0$. Il s'agit de la probabilité qu'une variable suivant une loi de Fisher soit au moins aussi grande que la valeur F_k observée. Par exemple, si p_k vaut 0.5, cela signifie qu'en tirant une variable selon $\mathcal{F}(1, N - d - 1)$, on a une chance sur deux d'observer un résultat au moins aussi grand que F_k. Dans ce cas, il est tout à fait plausible que $\beta_k = 0$, cette hypothèse n'est nullement contredite par la valeur F_k observée. Si au contraire p_k vaut 0.001, cela signifie qu'en tirant une variable selon $\mathcal{F}(1, N - d - 1)$, nous aurions une chance sur mille d'obtenir un résultat aussi grand que F_k. Dans ce cas, l'hypothèse $\beta_k = 0$ est très improbable.

En conclusion, la p-valeur quantifie notre croyance en le fait que le $k$ème régresseur contribue ou non à expliquer la réponse. Plus précisément, une p-valeur faible donne confiance en l'hypothèse $\beta_k \neq 0$, alors qu'une p-valeur de l'ordre de 0.5 donne confiance en l'hypothèse $\beta_k = 0$.

4.4. Cas général : estimateur du maximum de vraisemblance

Nous venons de voir deux modèles pour lesquels nous avons donné un estimateur raisonnable (et même optimal dans un certain sens dans le cas linéaire gaussien). Dans un cadre général, pouvons-nous définir des estimateurs raisonnables ?

4.4.1. Définition

On définit l'estimateur du *maximum de vraisemblance* (Maximum Likelihood -ML-, en anglais) comme suit

$$\hat{\boldsymbol{\Theta}}_{\mathrm{ML}}(y_1, \ldots, y_N) = \arg\max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(y_1, \ldots, y_N).$$

On peut avoir l'intuition que cet estimateur est raisonnable car il cherche le paramètre rendant l'observation disponible la plus vraisemblable.

Nous allons vérifier que cet estimateur nous permet de (re)-construire les estimateurs vus dans les deux modèles précédents et nous permet aussi d'en construire d'autres pour des modèles plus complexes.

4.4.2. Premiers exemples

Modèle Bernoulli

Nous avons vu via l'Eq. (4.1) que

$$p_\theta(y_n) = \theta^{y_n}(1-\theta)^{1-y_n}$$

avec $y_n \in \{0, 1\}$ pour $n = 1, \cdots, N$.

Si on suppose que la collection d'observations est iid, alors

$$p_\theta(y_1, \ldots, y_N) = \prod_{n=1}^{N} \theta^{y_n}(1-\theta)^{1-y_n} = \theta^{\sum_{n=1}^N y_n}(1-\theta)^{N-\sum_{n=1}^N y_n}$$

On a tracé sur la figure 4.4.2 la fonction $\theta \mapsto p_\theta(y_1, \ldots, y_N)$ avec $N = 30$, $\sum_{n=1}^N y_n = 10$. La fonction n'est pas concave sur $[0, 1]$ (qui est notre intervalle de recherche car le paramètre recherché est une probabilité) mais admet un unique maximum (une analyse formelle de son tableau de variation permettra de le montrer) et donc le maximum est atteint pour la dérivée nulle. Ainsi, en posant $s = \sum_{n=1}^N y_n$.

$$p'_\theta(\boldsymbol{y}) = \theta^s\theta^{N-s}.[s(1-\theta) - (N-s)\theta]$$

. Par conséquent, on obtient

$$\hat{\Theta}_{\mathrm{ML}}(y_1, \ldots, y_N) = \frac{1}{N}\sum_{n=1}^{N} y_n$$

ce qui est l'estimateur empirique de la Section 4.2.3.

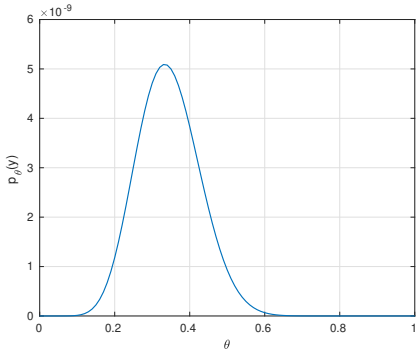


FIGURE 4.1. – Fonction $\theta \mapsto p_\theta(y_1, \ldots, y_N)$ dans le modèle de Bernoulli - $N = 30$, $\sum_n y_n = 10$. L'argument du maximum représente la valeur de θ la plus vraisemblable étant données les observations $y_1, \ldots, y_N$.

Modèle linéaire gaussien

Etant donné l'Eq. (4.5), il est facile de voir que l'estimateur des moindres carrés est en fait aussi l'estimateur du maximum de vraisemblance. Cela est laissé à titre d'exercice.

4.5. Régression logistique

4.5.1. Modèle paramétrique

Comme dans le cas de la régression linéaire, on se donne N couples

$$(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_N, Y_N)$$

où $\boldsymbol{x}_i \in \mathbb{R}^d$ est un vecteur de descripteurs, et Y_i est une variable de Bernoulli, à valeur dans $\{0, 1\}$. On fait l'hypothèse que l'on a accès à une réalisation $Y_1(\omega), \ldots, Y_N(\omega)$ des v.a. $Y_1, \ldots, Y_N$, qui correspondent à une issue d'une expérience aléatoire. La table B.1 fournit un exemple de jeu de données. A partir de ces réalisations, le statisticien souhaite proposer une *loi* des v.a. $Y_1, \ldots, Y_N$.

Obésité	Âge	Sexe
0	12	0
0	34	1
1	75	1
0	23	0
⋮		⋮
1	34	1

TABLE 4.4. – Détection d'une obésité au sein d'un groupe de N sujets. La réponse Y_i est égale à 1 si le sujet possède un indice de masse corporelle (imc) supérieur à 30, et vaut zéro sinon. Chaque descripteur est composé de deux variables, l'âge du sujet, et son sexe (0 = femme, 1= homme).

Cela revient à se fixer un modèle paramétrique, c'est à dire, faire une hypothèse sur la *forme* de

cette loi, en suivant le formalisme du paragraphe 4.1. Comme dans le cas de la régression linéaire, notre première hypothèse est :

Hypothèse 1 : Les observations $Y_1, \dots, Y_N$ sont indépendantes.

En conséquence, il suffit désormais de préciser les lois marginales de $Y_1, \dots, Y_N$. Pour tout $i = 1, \dots, N$, la loi de Y_i est nécessairement une loi de Bernoulli, puisque Y_i prend ses valeurs dans $\{0, 1\}$. Il nous reste seulement à préciser la probabilité que $\{Y_i = 1\}$. Nous faisons l'hypothèse que cette probabilité dépend d'un certain vecteur de paramètres $\boldsymbol{\theta}$, qu'il s'agira ensuite de déterminer à partir de nos observations. Dans le modèle de régression logistique, nous posons :

Hypothèse 2 : Il existe un vecteur $\boldsymbol{\theta} \in \mathbb{R}^{d+1}$ tel que pour tout i,
$$\mathbb{P}_{\boldsymbol{\theta}}(Y_i = 1) = \frac{1}{1 + \exp(-(\theta_0 + \theta_1 x_{i,1} + \cdots + \theta_d x_{i,d}))} .$$

Définition 4.6 (Fonction sigmoïde)**.** On définit la fonction sigmoïde $\varphi : \mathbb{R} \to (0, 1)$ par

$$\varphi(t) = \frac{1}{1 + e^{-t}}$$

pour tout $t \in \mathbb{R}$. C'est une fonction qui permet de transformer un nombre réel (souvent appelé un "score") en un nombre entre 0 et 1 (souvent appelé une "probabilité").

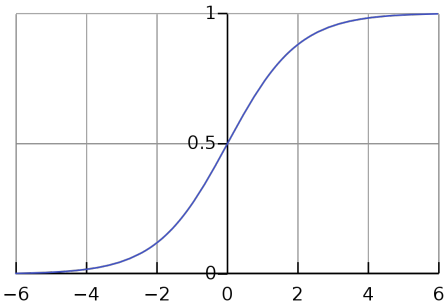


FIGURE 4.2. – Fonction sigmoide φ.

Discutons l'hypothèse 2. Nous supposons que la probabilité $\mathbb{P}_\theta(Y_i = 1)$ que la réponse soit égale à 1 est la fonction sigmoïde d'un *score*, qui est une fonctions affine des d attribus du $i$ème descripteur $\boldsymbol{x}_i$. Dans l'exemple de la table 4.4, cela revient à :

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Le sujet } i \text{ est obèse}) = \varphi(\theta_0 + \theta_1 \cdot \text{âge}_i + \theta_2 \cdot \text{sexe}_i) .$$

Encore une fois, on ne prétend jamais que cette expression est "vraie" pour un certain $\boldsymbol{\theta}$: il s'agit seulement d'un modèle, que nous jugeons pertinent pour expliquer les données. Pour obtenir une notation courte, on écrit le $i$ème score sous la forme d'un produit scalaire :

$$\theta_0 + \theta_1 x_{i,1} + \cdots + \theta_d x_{i,d} = \boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i$$

où $\tilde{\boldsymbol{x}}_i$ est le vecteur $\boldsymbol{x}_i$ augmenté d'un "1", soit :

$$\tilde{\boldsymbol{x}}_i = \begin{pmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,d} \end{pmatrix} .$$

Si le score est grand, la probabilité $\varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i)$ est proche de 1. Si le score est fortement négatif, elle est proche de 0. Si le score est proche de zéro, la probabilité est proche de 0.5.

On peut conclure ce paragraphe en rassemblant les hypothèses 1 et 2 précédentes. On résume le modèle paramétrique, tel que décrit dans le paragraphe 4.1. Nous rappelons les notations : $p_{\boldsymbol{\theta}}(k_1, \ldots, k_N) = \mathbb{P}_{\boldsymbol{\theta}}(Y_1 = k_1, \ldots, Y_N = k_N)$, et $p_{i,\boldsymbol{\theta}}(k) = \mathbb{P}_{\boldsymbol{\theta}}(Y_i = k)$. Nous avons donc $p_{i,\boldsymbol{\theta}}(1) = \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i)$. Par conséquent, $p_{i,\boldsymbol{\theta}}(0) = 1 - \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i)$. Nous pouvons rassembler ces deux expressions par :

$$\forall k \in \{0, 1\},\ p_{i,\boldsymbol{\theta}}(k) = \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i)^k (1 - \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i))^{1-k} .$$

On a finalement :

Modèle de régression logistique. Pour tout $(k_1, \ldots, k_N) \in \{0, 1\}^N$,

$$\begin{aligned} p_{\boldsymbol{\theta}}(k_1, \ldots, k_N) &= \prod_{i=1}^{N} p_{i,\boldsymbol{\theta}}(k_i) \\ &= \prod_{i=1}^{N} \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i)^{k_i} (1 - \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i))^{1-k_i} . \end{aligned}$$

4.5.2. Maximum de vraisemblance

Une fois le modèle paramétrique fixé, nous pouvons exprimer l'estimateur du maximum de vraisemblance. Il s'agit de la fonction $\hat{\Theta}_{\mathrm{ML}}$ qui à toute observation possible $(k_1, \ldots, k_N) \in \{0, 1\}^N$ associe la valeur de $\boldsymbol{\theta}$ maximisant la vraisemblance $p_{\boldsymbol{\theta}}(k_1, \ldots, k_N)$. Pour des raisons numériques, on préfère souvent maximiser des sommes plutôt que des produits. On préfèrera donc maximiser la *log-vraisemblance* $\ln p_{\boldsymbol{\theta}}(k_1, \ldots, k_N)$. Au final, on a l'expression :

$$\hat{\Theta}_{\mathrm{ML}}(k_1, \ldots, k_N) = \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \sum_{i=1}^{N} \left(k_i \ln \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i) + (1 - k_i) \ln(1 - \varphi(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}_i)) \right) .$$

C'est à dire que si on observe $Y_1, \ldots, Y_N$ représentent les v.a. observées, l'estimée est :

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \hat{\Theta}_{\mathrm{ML}}(Y_1, \ldots, Y_N) .$$

Il n'existe malheureusement pas de forme explicite pour la solution de ce problème de minimisation. En pratique, il faut recourrir à une procédure itérative, appelée un algorithme d'optimisation numérique, pour obtenir une approximation de l'estimée. Par exemple, un algorithme du gradient, ou encore un algorithme de Newton-Raphson.

4.5.3. Prédiction

Une fois l'estimée $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ au sens du maximum de vraisemblance obtenue, la question est de savoir comment prédire la réponse Y associée à une nouvelle entrée $\boldsymbol{x}$, non présente dans notre N-échantillon. La prédiction peut prendre la forme d'une *décision douce* :

— Le score associé à $\boldsymbol{x}$ est $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^T \tilde{\boldsymbol{x}}$ (où $\tilde{\boldsymbol{x}}$ est le vecteur $\boldsymbol{x}$ augmenté d'un "1") ;

— La probabilité associée à $\boldsymbol{x}$ est $\varphi(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^T \tilde{\boldsymbol{x}})$. Ce nombre entre 0 et 1 représente notre croyance en l'événement $\{Y = 1\}$, connaissant $\boldsymbol{x}$. Cette croyance dépend de notre jeu de donnés au travers de l'estimateur $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$.

Notons, $\hat{h}(\boldsymbol{x})$ la fonction qui à tout $\boldsymbol{x}$ associe la décision douce (probabilité ou score, comme on veut) :

$$\hat{h}(\boldsymbol{x}) = \varphi(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^T \tilde{\boldsymbol{x}}) \,.$$

La décision douce est plus informative qu'une prédiction de type "0" ou "1", car elle quantifie la croyance en le fait que la réponse est 0 ou 1.

Si malgré cela il s'agit vraiment de prédire la valeur 0 ou 1 pour la réponse, l'usage est de comparer décision douce à un seuil, et de décider $\hat{Y} = 1$ si le résultat est supérieur au seuil, et $\hat{Y} = 0$ sinon. On parle alors de *décision dure*. Notons $\hat{h}_{\mathrm{dur}}(\boldsymbol{x})$ la décision dure associée à une entrée $\boldsymbol{x}$. On a :

$$\hat{h}_{\mathrm{dur}}(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \hat{h}(\boldsymbol{x}) > \tau \\ 0 & \text{sinon.} \end{cases}$$

Ici, le réel τ est le seuil de décision. Si la décision douce est une probabilité, ce seuil doit être choisi dans l'intervalle $(0, 1)$. Si la décision douce est un score, il doit être choisi dans $\mathbb{R}$.

Le choix "par défaut" pour un seuil de probabilité est $\tau = 0.5$. Il correspond à décider "1" si la probabilité estimée que $Y = 1$ connaissant $\boldsymbol{x}$ est supérieure à 0.5. Mais en réalité, le choix du seuil dépend de l'application.

4.5.4. Borne de Cramer-Rao

Note PB : Je suggère de maintenir dans ce chapitre la convention :

— $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ la variable aléatoire ;

— $\boldsymbol{Y}(\omega) = (Y_1(\omega) \ldots, Y_N(\omega))^T$ la réalisation qui est effectivement observée par le statisticien lors d'une certaine expérience aléatoire ;

— $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ qui est une variable muette, un point de $\mathbb{R}^N$.

Dans cette section, on va montrer que pour une certaine classe d'estimateurs, l'erreur quadratique moyenne de tout estimateur de cette classe est bornée inférieurement par une certaine valeur, appelée *borne de Cramer-Rao*, que l'on est capable de caractériser et parfois de calculer analytiquement.

4.5.5. Cas simple

On considère l'ensemble des estimateurs $\hat{\theta}$ non biaisé de θ (avec θ un paramètre scalaire à estimer). Ce cas simple est aussi dit cas scalaire ou cas mono-varié.

Sous certaines conditions techniques sur la fonction de vraisemblance (qu'on introduira au cours de la démonstration), on a que pour tout $\hat{\theta}$ non biaisé,

$$\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \mathrm{BCR}(\theta)$$

avec

— la BCR qui est donnée par

$$\mathrm{BCR}(\theta) = \frac{1}{F(\theta)}$$

où $F(\theta)$ est l'information de Fisher décrite par

$$F(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial \theta} \right)^2 \right]$$

et $p_\theta(\boldsymbol{y})$ est la vraisemblance des données $\boldsymbol{y}$ paramétrée par θ.

— L'espérance mathématique est à prendre sur toutes les variables aléatoires du problème.

Démonstration. On pose

$$\begin{aligned}
\mathbb{E}_\theta \left[(\hat{\theta} - \theta).\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial \theta} \right] &= \int (\hat{\theta} - \theta).\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial \theta} p_\theta(\boldsymbol{y}) d\boldsymbol{y} \\
&\overset{(a)}{=} \int (\hat{\theta} - \theta).\frac{\partial p_\theta(\boldsymbol{y})}{\partial \theta} d\boldsymbol{y} \\
&= \int \hat{\theta}\frac{\partial p_\theta(\boldsymbol{y})}{\partial \theta} d\boldsymbol{y} - \int \theta\frac{\partial p_\theta(\boldsymbol{y})}{\partial \theta} d\boldsymbol{y} \\
&\overset{(b)}{=} \frac{\partial}{\partial \theta} \int \hat{\theta} p_\theta(\boldsymbol{y}) d\boldsymbol{y} - \theta\frac{\partial}{\partial \theta} \int p_\theta(\boldsymbol{y}) d\boldsymbol{y} \\
&\overset{(c)}{=} \frac{\partial}{\partial \theta}\mathbb{E}_\theta[\hat{\theta}] - \theta\frac{\partial 1}{\partial \theta} \\
&\overset{(d)}{=} \frac{\partial \theta}{\partial \theta} \\
&= 1
\end{aligned}$$

L'égalité (a) vient de la propriété de dérivation du ln. Pour le premier terme dans (b), on utilise le fait que $\hat{\theta}$ ne dépend que des données (et non de θ) et qu'on suppose qu'on peut sortir la dérivation de l'intégrale. Pour la seconde partie de (b), on sort le θ de l'intégrale car on intègre sur $\boldsymbol{y}$ et de nouveau on suppose que la dérivation peut être sortie de l'intégrale. Pour (c), on utilise la définition de l'espérance (et ceci marche car le p_θ est calculé sur la vraie valeur de paramètre) et on utilise aussi le fait qu'une densité de probabilité intègre à 1. Enfin (d) est obtenu grâce à l'hypothèse de l'estimateur non-biaisé (l'hypothèse forte ne sert que là !).

Et en appliquant Cauchy-Schwarz, on a bien

$$\mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right] . \mathbb{E}_\theta \left[\left(\frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial \theta} \right)^2 \right] \geq 1$$

ce qui conclut la preuve.□

Exemple 4.7. On considère le cas gaussien linéaire simple. On a un N-échantillon avec

$$y_i = \theta x_i + \varepsilon_i$$

avec ε_i iid gaussien de moyenne nulle et de variance σ^2 connue.

On calcule donc la log-vraisemblance. On a

$$\begin{aligned} \ln p_\theta(\boldsymbol{y}) &= \ln \prod_{i=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - x_i\theta)^2}{2\sigma^2}} \right) \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{N} (y_i - x_i\theta)^2}{2\sigma^2}. \end{aligned}$$

En dérivant, on a

$$\begin{aligned} \frac{\partial \ln p_\theta(\boldsymbol{y})}{\partial \theta} &= \frac{\sum_{i=1}^{N} (y_i - x_i\theta) x_i}{\sigma^2} \\ &\overset{(a)}{=} \frac{\sum_{i=1}^{N} \varepsilon_i x_i}{\sigma^2}. \end{aligned}$$

Dans (a), on remplace y_i par son modèle.

Pour l'information de Fisher, on a

$$\begin{aligned} F(\theta) &= \mathbb{E}_\theta \left[\left(\frac{\sum_{i=1}^{N} \varepsilon_i x_i}{\sigma^2}. \right)^2 \right] \\ &= \frac{1}{\sigma^4} \sum_{i,j=1}^{N} x_i x_j \mathbb{E}[\varepsilon_i \varepsilon_j] \\ &\overset{(a)}{=} \frac{1}{\sigma^4} \sum_{i=1}^{N} x_i^2 \mathbb{E}[\varepsilon_i^2] \\ &= \frac{\sum_{i=1}^{N} x_i^2}{\sigma^2}. \end{aligned}$$

L'égalité (a) est obtenue en observant que $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$ pour $i \neq j$.

Finalement cela conduit à

$$\mathrm{BCR}(\theta) = \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2}.$$

On voit que la BCR diminue quand le bruit diminue et aussi quand la taille du N-échantillon augmente. Mais pouvez-vous dire comment il évolue en fonction de N si $x_i \in \{-1, 1\}$? Pensez-vous que cela est un comportement généralisable ?

4.5.6. Cas multiple

On considère maintenant un paramètre vectoriel $\boldsymbol{\theta}$. Ce cas est dit aussi multi-varié.

Toujours sous certaines conditions techniques (que l'on précisera au cours de la preuve), on a que la matrice des erreurs quadratiques est inférieure (au sens de l'ordre partiel des matrices semi-définies positives) à l'inverse de la matrice d'information de Fisher. Ainsi

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T\right] \succeq \mathbf{F}(\boldsymbol{\theta})^{-1} \tag{4.7}$$

où

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ln p_{\boldsymbol{\theta}}(\boldsymbol{y})}{\partial \boldsymbol{\theta}}\right)^T\right].$$

La matrice $\mathbf{F}$ est dite matrice d'information de Fisher et son inverse est dite matrice de la borne de Cramer-Rao.

Démonstration : en classe.

Notez que l'erreur quadratique est obtenue en appliquant la trace de part et d'autre de l'Eq. (4.7) et l'erreur quadratique de chaque composante de $\boldsymbol{\theta}$ en prenant le terme diagonal correspondant de la matrice de BCR. Ainsi, on a

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(\hat{\theta}_d-\theta_d)^2\right] \geq \left[\mathbf{F}(\theta)^{-1}\right]_{d,d} \tag{4.8}$$

et

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\right] \geq \text{trace}\left(\mathbf{F}(\theta)^{-1}\right). \tag{4.9}$$

Exemple 4.8. Prenons le cas gaussien avec la moyenne à calculer et la variance à calculer. Ainsi, on considère que

$$y_i = m + \varepsilon_i$$

avec ε_i un processus iid gaussien de moyenne nulle et de variance inconnue σ^2.

On a donc $\boldsymbol{\theta} = [m, \sigma^2]^T$.

On obtient que

$$\mathbf{F}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}.$$

Essayez de trouver le résultat suivant par vous-même et en déduire la BCR pour chaque composante de $\boldsymbol{\theta}$.

4.5.7. Sélection de modèle

Dans le cadre de la sélection de modèle, on peut aussi utiliser le principe de maximum de vraisemblance.

Imaginons de nouveau un N-échantillon dont nous cherchons un modèle approprié entre les y_i et les $\boldsymbol{x}_i$ avec une taille D pour les vecteurs $\boldsymbol{x}_i$. On appelle $\mathcal{M}_D$ le modèle dépendant de D variables explicatives dépendant donc d'un paramètre multi-varié $\boldsymbol{\theta}$ de taille D.

La vraisemblance des $\boldsymbol{y}$ dépend donc du modèle choisi et du paramètre associé. On la notera

$$p_{\mathcal{M}_D,\boldsymbol{\theta}}(\boldsymbol{y}).$$

La vraisemblance du modèle sera obtenue en remplaçant le paramètre $\boldsymbol{\theta}$ inconnu par un de ces estimateurs, typiquement celui du maximum de vraisemblance. Par conséquent la vraisemblance des données par rapport au modèle $\mathcal{M}_D$ est noté et vaut respectivement

$$p_{\mathcal{M}_D}(\boldsymbol{y}) = p_{\mathcal{M}_D,\hat{\boldsymbol{\theta}}_{\mathrm{ML}}}(\boldsymbol{y}).$$

Néanmoins si les modèles $\{\mathcal{M}_D\}_{D\in\mathcal{D}}$ sont emboîtés ce qui signifie qu'en forçant une composante de $\boldsymbol{\theta}$ à zéro dans le modèle $\mathcal{M}_D$, on obtient le modèle $\mathcal{M}_{D-1}$, il est clair qu'on choisira le modèle avec le plus de paramètres ce qui n'est pas ce qu'on souhaite non plus car on veut conserver un modèle simple. C'est pourquoi, on va pénaliser les modèles ayant trop de paramètres via la fonction $\mathrm{Pen}(\mathcal{M}_D)$. Ainsi

$$\hat{\mathcal{M}} = \arg \min_{\{\mathcal{M}_D\}_{D\in\mathcal{D}}} -\ln\left(p_{\mathcal{M}_D}(\boldsymbol{y})\right) + \mathrm{Pen}(\mathcal{M}_D).$$

avec, par exemple,

— $\mathrm{Pen}(\mathcal{M}_D) = D$ pour le critère d'information d'Akaike (AIC)

— $\mathrm{Pen}(\mathcal{M}_D) = \ln(N).D$ pour le critère d'information bayésienne (BIC)

4.6. Exercices

Exercice 4.1. En utilisant l'inégalité de Chebychev-Cantelli (2.9) (voir l'exercice 2.10), fournir un intervalle de confiance pour le modèle de Bernoulli. Est-il plus exact que celui de l'équation (4.2) ?

Exercice 4.2. L'*inégalité de Hoeffding*[1] stipule que, si $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes, telles que pour tout i, $a_i \leq X_i \leq b_i$. Soit $S_n = X_1 + \cdots + X_n$. Alors pour tout $\epsilon > 0$,

$$\mathbb{P}\left(S_n - \mathbb{E}(S_n) > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

En utilisant cette inégalité, fournir un intervalle de confiance pour le modèle de Bernoulli. Est-il plus exact que celui de l'équation (4.2) ?

Exercice 4.3. On considère des variables aléatoires Y_i $(i = 1 \ldots N)$ sur $\mathbb{R}$ suivant un modèle homoscédastique :

$$Y_i = \langle \boldsymbol{\beta}, \boldsymbol{x}_i \rangle + \varepsilon_i \,,$$

où pour tout i, $\boldsymbol{x}_i \in \mathbb{R}^d$ est un vecteur colonne décrivant les variables explicatives du $i$ème exemple, $\boldsymbol{\beta}$ est un vecteur déterministe inconnu de $\mathbb{R}^d$, et où les ε_i sont des v.a. iid gaussiennes centrées de variance σ^2.

1. Rappeler l'expression de l'estimateur $\hat{\boldsymbol{\beta}}$ au sens des moindres carrés.

1. Ceux que cela intéresse iront visiter la page Wikipedia correspondante, pour la preuve de cette inégalité.

2. On considère un nouveau couple $(\boldsymbol{x}, Y)$ qui ne figure pas dans la base de données. La v.a. Y est non-observée. Exprimer la prédiction $\hat{Y}$ de Y à partir de l'estimateur des moindres carrés.
3. Exprimer $\hat{\boldsymbol{\beta}}$ en fonction de $\boldsymbol{\beta}$ et du vecteur $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^T$.
4. On suppose que $(\boldsymbol{x}, Y)$ suit le même modèle homoscédastique :

$$Y = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle + \varepsilon \,,$$

   Exprimer $\hat{Y}$ en fonction de Y, $\boldsymbol{x}$, ε et $\boldsymbol{\varepsilon}$.
5. La prédiction est-elle biaisée ?
6. Caractériser la loi de $Y - \hat{Y}$.
7. En supposant σ^2 connue, fournir un intervalle de confiance sur Y.
8. Que peut-on faire si σ^2 est inconnu ? Proposer une alternative.

Exercice 4.4. On se place dans le même contexte que l'exercice précédent (mêmes notations). On résoud cette fois le problème des moindres carrés régularisés :

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \,,$$

où $\lambda \geq 0$ est un certain hyper-paramètre.

1. Fournir la solution explicite.
2. Exprimer $\hat{\boldsymbol{\beta}}_\lambda$ en fonction de $\boldsymbol{\beta}$. Cet estimateur est-il biaisé ?
3. Calculer l'espérance de l'erreur quadratique commise sur le nouvel échantillon $\mathbb{E}((Y - \boldsymbol{x}^T \hat{\boldsymbol{\beta}}_\lambda)^2)$. Discuter de l'influence de λ. Quel intérêt peut-il y avoir à choisir $\lambda \neq 0$.
4. Ecrire une procédure, utilisable en pratique à partir d'un jeu de données $(\boldsymbol{x}_i, Y_i)$ pour $i = 1, \dots, N$, et permettant de choisir un λ convenable.

Exercice 4.5. On cherche à estimer l'espérance de vie d'un patient après diagnostic d'une certaine maladie. Pour cela, on appelle T la durée de vie après diagnostic d'un patient aléatoire, mesurée en mois (on pose $T = 1$ si le patient décède durant le premier mois, $T = 2$ s'il décède durant le deuxième, etc.). On se donne un modèle paramétrique : à chaque nouveau mois, le patient a une probabilité $\theta \in (0, 1)$ de décès, indépendante des mois précédents. Ici, θ est le paramètre inconnu de notre modèle paramétrique.

1. Donner la loi de T en fonction de θ, c'est à dire, calculer $\mathbb{P}_\theta(T = k)$ pour tout k. Reconnaître cette loi.
2. Calculer l'espérance de T en fonction de θ.
3. On suppose que l'on observe les durées de vie $T_1, \dots, T_n$ de n patients. Calculer l'estimée de θ au sens du maximum de vraisemblance.
4. En déduire un estimateur de l'espérance de survie après diagnostic, à partir des observations $T_1, \dots, T_n$.
5. A partir des observations $T_1, \dots, T_n$, estimer la probabilité de survivre au moins deux ans après le diagnostic.

6. On effectue une expérimentation sur une durée limitée. Entre le début de l'expérimentation à la date d_0 et la fin de l'expérimentation à la date d_1, n patients sont diagnostiqués. Dans ce laps de temps $[d_0, d_1]$, certains patients meurent, et d'autres sont toujours vivants à la fin de l'expérimentation. Pour les patients i survivants, on n'observe pas la durée de vie T_i, on observe seulement l'événement que T_i est plus grand que la durée τ_i qui sépare l'instant du diagnostic du patient i de la fin de l'expérimentation. Autrement dit, on observe la variable :
$$Y_i = \begin{cases} T_i & \text{si } T_i \leq \tau_i \\ \infty & \text{si } T_i > \tau_i \,, \end{cases}$$
A partir des seules observations $Y_1, \ldots, Y_n$, estimer l'espérance de survie après diagnostic.

On sait que, en l'absence de traitement, la durée moyenne de survie après diagnostic est égale à $D = 12$ mois. On teste un nouveau traitement, et on cherche à savoir si ce traitement possède une influence significative sur la durée de vie. On suppose à nouveau que l'on traite n patients avec ce nouveau traitement, et que l'on dispose des observations $T_1, \ldots, T_n$ correspondant à leurs durées de vie après diagnostic. On supposera toujours que T_i suit une loi géométrique de paramètre θ inconnu.

7. Montrer que le problème revient à tester l'hypothèse $\theta \neq 1/D$ (dite l'alternative) contre l'hypothèse $\theta = 1/D$ (dite hypothèse nulle).
8. On rappelle/admet que la variance d'une loi géométrique de paramètre θ est égale à $(1 - \theta)/\theta^2$. Sous l'hypothèse nulle, calculer l'espérance et la variance de la variable aléatoire :
$$S = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{T_i - D}{\sqrt{D(D-1)}}$$
9. En utilisant le théorème central limite, montrer que, si n est assez grand, on peut supposer que S^2 suit une loi du chi-deux à un degré de liberté (on rappelle qu'une loi du chi deux à un degré de liberté est la loi suivie par le carré d'une v.a. gaussienne centrée réduite).
10. Regarder sur Wikipedia (ou autre) l'allure de la loi du chi-deux à un degré de liberté. Dans notre expérimentation, on observe que $S^2 = 0.93$. Peut-on conclure à un effet significatif du nouveau traitement ? Même question pour $S^2 = 9.21$. Justifier votre raisonnement.
11. On désigne par $\bar{F}(x)$ la fonction de répartition complémentaire de la loi du chi deux à un degré de liberté, c'est à dire $\bar{F}(x) = \mathbb{P}(Z^2 > x)$, où Z^2 suit une loi du chi deux à un degré de liberté. La quantité $p = \bar{F}(S^2)$ est appelé la p-valeur associée au test d'hypothèse. Discuter de l'interprétation de la p-valeur : que signifie un p-valeur de l'ordre de 0.5 ? que signifie une p-valeur de l'ordre de 0.05 ? Dans quel cas jugeriez vous que le nouveau traitement a un impact significatif sur la durée de vie ?

Exercice 4.6. Soit $X_1, \ldots, X_n$ des v.a. iid de loi uniforme sur l'intervalle $[0, \theta]$. Calculer l'estimateur de θ au sens du macimum de vraisemblance.

Exercice 4.7. On souhaite estimer une fréquence f_0 dépendant de cette manière des observations
$$y_n = e^{2i\pi f_0 n} + \varepsilon(n)$$
avec i le nombre complexe tel que $i^2 = -1$ et $\varepsilon(n)$ un processus gaussien iid de moyenne nulle et de variance σ^2 (et de parties réelles et imaginaires indépendantes entre elles).

Donner la formule de l'estimateur du maximum de vraisemblance pour f_0. Obtient-on un résultat intuitif ?

A. Eléments d'analyse convexe

Si $f : \mathbb{R}^d \to \mathbb{R}$ est une fonction, on dit que x est un *minimiseur* de f si $f(y) \geq f(x)$ pour tout y. On note $\arg\min f$ l'ensemble des minimiseurs de f (c'est un ensemble, mais lorsque f admet un unique minimiseur, $\arg\min f$ est simplement un point de $\mathbb{R}^d$). Le gradient d'une fonction f en un point $x \in \mathbb{R}^d$ est le vecteur :

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{pmatrix} .$$

Dans le cas unidimensionnel $d = 1$, le gradient est simplement la dérivée $f'(x)$ de la fonction. On peut aller un cran plus loin, et définir la *matrice hessienne* de f au point x. Il s'agit de la matrice $d \times d$ définie par :

$$\mathrm{Hess}(f) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_d} \\ \vdots & & & \vdots \\ \frac{\partial^2 f(x)}{\partial x_d \partial x_1} & \frac{\partial^2 f(x)}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_d^2} \end{pmatrix}$$

c'est à dire que le coefficient (i, j) de la matrice est $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. Cette matrice est symétrique, car on peut permuter l'ordre de dérivation entre x_i et x_j. Dans le cas unidimensionnel $d = 1$, le Hessien est simplement la dérivée-seconde $f''(x)$ de la fonction.

Un point x est un *point critique* de f si $\nabla f(x) = 0$. Tout minimiseur de f est un point critique. La réciproque n'est pas vraie en générale, mais elle est vraie si f est une fonction convexe. Une fonction f est dite *convexe* si la propriété suivante est satisfaite :

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

pour tout $t \in [0, 1]$ et tous points x et y dans $\mathbb{R}^d$. On notera que l'ensemble $\{tf(x)+(1-t)f(y), t \in [0, 1]\}$ est le segment de droite reliant $f(x)$ à $f(y)$.

Théorème A.1. *Dans le cas où f est convexe, on a effectivement l'équivalence :*

$$x \in \arg\min f \;\Leftrightarrow\; \nabla f(x) = 0 \,.$$

Cela signifie que pour trouver un minimiseur d'une fonction il suffit de chercher un point qui annule le gradient. De ce point de vue, les fonctions convexes sont sympathiques, car on a alors un outil simple pour rechercher ces minimiseurs. En pratique, il faudra savoir repérer qu'une

fonction est convexe. Dans le cas unidimensionnel $d = 1$ (f est une fonction de $\mathbb{R} \to \mathbb{R}$), on se convainc facilement que f est convexe si et seulement sa dérivée est croissante (faire un dessin pour s'en convaincre). Autrement dit, f est convexe si et seulement si sa dérivée-seconde est partout positive ou nulle :

$$f : \mathbb{R} \to \mathbb{R} \text{ est convexe } \Leftrightarrow \ \forall x \in \mathbb{R},\ f''(x) \geq 0\,.$$

Ce résultat intuitif admet une généralisation dans $\mathbb{R}^d$, donnée par le théorème suivant :

Théorème A.2. *Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est convexe si et seulement si* $\mathrm{Hess}(f)$ *est semi-définie positive.*

B. Tests d'hypothèses

Note PB : Je suggère de maintenir dans ce chapitre la convention :

- $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$ la variable aléatoire ;
- $\boldsymbol{Y}(\omega) = (Y_1(\omega) \ldots, Y_N(\omega))^T$ la réalisation qui est effectivement observée par le statisticien lors d'une certaine expérience aléatoire ;
- $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ qui est une variable muette, un point de $\mathbb{R}^N$.

B.1. Introduction

Dans nombreuses situations pratiques, on aimerait savoir si une hypothèse est plus juste qu'une autre. On l'a vu dans le chapitre précédent avec un test de dépendance d'un modèle par rapport à une variable explicative ou bien dans la sélection de modèle (dans le cas alors de deux modèles). Historiquement, ce problème s'est d'abord rencontré dans des contextes militaires où une hypothèse (dite $\mathcal{H}_1$)) correspond à la détection d'un missile ou d'un avion et l'autre hypothèse (dite $\mathcal{H}_0$) correspond à l'espace sûr. Dans ce dernier exemple, on voit clairement que les deux hypothèses ne sont pas mises au même niveau ce qui induira une différence d'analyse de performance entre les deux.

D'abord, on considère que la loi des données $\boldsymbol{y}$ est indexée par l'hypothèse et est donc différente d'une hypothèse à l'autre. Ainsi on a

$$\begin{cases} \text{Hypothèse } \mathcal{H}_0 : & \boldsymbol{y} & \sim & p_{\mathcal{H}_0}(\boldsymbol{y}) \\ \text{Hypothèse } \mathcal{H}_1 : & \boldsymbol{y} & \sim & p_{\mathcal{H}_1}(\boldsymbol{y}) \end{cases}$$

On notera $\hat{\mathcal{H}}$ l'hypothèse détectée dans toute la suite.

On peut définir trois types usuels de performance comme dans le Tableau suivant.

Hyp. détectée \ Vraie hyp.	$\mathcal{H}_0$	$\mathcal{H}_1$
$\hat{\mathcal{H}} = \mathcal{H}_0$	$\times$	$P_M = \Pr_{\mathcal{H}_1}\{\mathcal{H}_0\}$ Probabilité de mauvaise détection Probabilité d'erreur de type II
$\hat{\mathcal{H}} = \mathcal{H}_1$	$P_{FA} := \Pr_{\mathcal{H}_0}\{\mathcal{H}_1\}$ Probabilité de fausse alarme Probabilité d'erreur de type I	$P_D = \Pr_{\mathcal{H}_1}\{\mathcal{H}_1\} = 1 - P_M$ Puissance de bonne détection Puissance du test

TABLE B.1. – Différentes types de performance et leurs relations

Dans le cadre historique, il apparaît raisonnable de vouloir maximiser P_D ce qui est très facile à faire en choisissant toujours $\mathcal{H}_1$ et donc il faut interdire ce choix de test en bornant aussi la P_{FA} (qui serait aussi égale à 1 avec le test trivial précédent). Ceci va conduire au test suivant décrit dans la prochaine section.

B.2. Test optimal

Théorème B.1 (Test de Neyman-Pearson)**.** *Maximiser la probabilité de bonne détection sous la condition que la probabilité de fausse alarme soit en-dessous d'un certain niveau (noté P_{FA}^t) conduit au test de Neyman-Pearson suivant, également, appelé test du Rapport de Vraisemblance (Likelihood Ratio Test-LRT , en anglais),*

$$T(\boldsymbol{y}) = \ln\left(\frac{p_{\mathcal{H}_1}(\boldsymbol{y})}{p_{\mathcal{H}_0}(\boldsymbol{y})}\right) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \mu,$$

avec

- *T le Log Likelihood Ratio-LLR*
- *μ le seuil du test qui permet de satisfaire le niveau-cible de probabilité de fausse alarme.*

Démonstration : en classe.

Evidemment pour mettre en place ce test en pratique, il faut être capable de trouver une forme analytique pour $p_{\mathcal{H}_0}$ et $p_{\mathcal{H}_1}$ et idéalement d'ajuster le seuil μ aussi mathématiquement.

Noter que le test trivial décrit à la section précédente s'écrit $T(\boldsymbol{y}) = 1, \forall \boldsymbol{y}$, et alors $P_D = 1$ et $P_{FA} = 1$ ce qui évidemment dans un contexte militaire est absurde car cela reviendrait à détecter un missile ou un avion ennemi à tous les coups et réduire les stocks à une vitesse grand V.

Il est clair aussi que pour un test non-trivial, P_D va fortement dépendre de P_{FA} notamment en jouant sur le seuil μ. Par conséquent, il serait intéressant de tracer P_D en fonction de P_{FA}.

Définition B.2. Pour une configuration donnée (Rapport Signal-à-Bruit donné, nombre d'échantillons N donné, etc), la fonction $P_{FA} \mapsto P_D$ est appelée courbe *Receiver Operating Characteristics-ROC.*

Comment la trace-t-on ? Typiquement comme une courbe paramétrée par μ en reliant les points $(P_{FA}(\mu), P_D(\mu))$ dans l'ordre des μ.

Exemple B.3.

$$\begin{cases} \mathcal{H}_0 & : y_n = w_n \\ \mathcal{H}_1 & : y_n = x_n + w_n \end{cases}, n = 1, \cdots, N$$

avec

- w_n suite gaussienne iid de moyenne nulle et de variance (connue) $\sigma_w^2 = \mathbb{E}[w_n^2]$,

— x_n également une suite gaussienne iid de moyenne nulle et de variance (connue) $\sigma_x^2 = \mathbb{E}[x_n^2]$.
On a

$$\left\{ \begin{array}{rcl} p_{\mathcal{H}_0}(\boldsymbol{y}) & = & \prod_{n=1}^N p_{\mathcal{H}_0}(y_n) \text{ avec } p_{\mathcal{H}_0}(y_n) = \frac{1}{(2\pi\sigma_w^2)^{1/2}} e^{-\frac{y_n^2}{2\sigma_w^2}} \\ p_{\mathcal{H}_1}(\boldsymbol{y}) & = & \prod_{n=1}^N p_{\mathcal{H}_1}(y_n) \text{ avec } p_{\mathcal{H}_1}(y_n) = \frac{1}{(2\pi(\sigma_x^2+\sigma_w^2))^{1/2}} e^{-\frac{y_n^2}{2(\sigma_x^2+\sigma_w^2)}} \end{array} \right. .$$

d'où

$$\begin{aligned} T(\boldsymbol{y}) &= \ln \left(\frac{\frac{1}{(2\pi)^{N/2}(\sigma_x^2+\sigma_w^2)^{N/2}} e^{-\frac{\sum_{n=1}^N y_n^2}{2(\sigma_x^2+\sigma_w^2)}}}{\frac{1}{(2\pi)^{N/2}\sigma_w^N} e^{-\frac{\sum_{n=1}^N y_n^2}{2\sigma_w^2}}} \right) \\ &= \ln \left(\left(\frac{\sigma_w^2}{\sigma_x^2+\sigma_w^2} \right)^{N/2} e^{-(\frac{1}{2(\sigma_x^2+\sigma_w^2)} - \frac{1}{2\sigma_w^2}) \sum_{n=1}^N y_n^2} \right) \\ &= \text{constante positive} \times \sum_{n=1}^N y_n^2 + \text{constante} \end{aligned}$$

Le test LRT est donc un test d'énergie. On peut en fait choisir les constantes à notre guise. Et donc on fera en sorte que le test final soit le suivant.

$$T(\boldsymbol{y}) = \frac{1}{\sigma_x^2+\sigma_w^2} \sum_{n=1}^N y_n^2 \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta.$$

— Sous $\mathcal{H}_1$, $T(\boldsymbol{y})$ suit une loi du χ_2 à N degrés de liberté

$$p_{\chi_2,N}(x) = \frac{1}{\Gamma_c(N/2)} x^{N/2-1} e^{-x}, \ x \geq 0$$

— Sous $\mathcal{H}_0$, $T(\boldsymbol{y})$ suit une loi du χ_2 à N degrés de liberté

$$p_{\chi_2,N}(x) = \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^{N/2}\Gamma_c(N/2)} x^{N/2-1} e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}}, \ x \geq 0$$

avec les fonctions Gamma complète et incomplète suivante

$$\Gamma_c(s) = \int_0^\infty x^{s-1} e^{-x} dx$$

et

$$\Gamma_{\text{inc}}(s,u) = \int_u^\infty x^{s-1} e^{-x} dx.$$

Par conséquent, on obtient

$$\begin{aligned} P_{FA} &= \Pr_{\mathcal{H}_0}(T(\boldsymbol{y}) > \eta) \\ &= \int_\eta^\infty \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^{N/2}\Gamma_c(N/2)} x^{N/2-1} e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}} dx \\ &= \frac{1}{\Gamma_c(N/2)} \cdot \frac{1}{(\sigma_w^2/(\sigma_x^2+\sigma_w^2))^{N/2}} \cdot \int_\eta^\infty x^{N/2-1} e^{-\frac{(\sigma_x^2+\sigma_w^2)x}{\sigma_w^2}} dx \\ &= \frac{\Gamma_{\text{inc}}\left(N/2, \eta\frac{\sigma_x^2+\sigma_w^2}{\sigma_w^2}\right)}{\Gamma_c(N/2)}. \end{aligned}$$

De manière similaire, on a

$$P_D = \frac{\Gamma_{\text{inc}}(N/2, \eta)}{\Gamma_c(N/2)}.$$

Sur la figure B.1, on trace pour le test précédent la courbe ROC. On souhaite évidemment que la courbe se rapproche le plus rapidement possible du coin Nord-Ouest.

FIGURE B.1. – Courbe ROC pour le test d'hypothèse LRT (avec RSB= -20dB)

B.3. Lien entre intervalle de confiance et test

On peut voir un intervalle de confiance comme un test qu'un certain paramètre θ recherché appartienne à cet intervalle avec une certaine probabilité.

Plus précisément, quand on veut savoir si le paramètre recherché vaut 0 (et donc sera associé à une variable non explicative), on associera cette hypothèse à $\mathcal{H}_0$. Alors le test détectera correctement $\mathcal{H}_0$ (avec une certaine probabilité qui sera alors la confiance $(1-\alpha)$ que l'on a sur le fait que la variable est non explicative) si le test $T(\boldsymbol{y})$ est inférieur à un seuil. Ce test peut être par exemple un estimateur de θ que l'on note $\hat{\theta}(\boldsymbol{y})$. Par conséquent

$$1 - P_{FA} = \Pr_{\mathcal{H}_0}(\hat{\theta}(\boldsymbol{y}) \leq \mu) = 1 - \alpha.$$

B.4. Exercices

Exercice B.1.

- On considère des échantillons bi-dimensionnel (bref, $y \in \mathbb{R}^2$).
- On considère que l'hypothèse 0 correspond à une gaussienne (bi-dimensionnelle) de moyenne $\boldsymbol{m}_0$ et de variance σ_0^2 (chaque composante de y est indépendante et de même variance σ_0^2).
- On considère que l'hypothèse 1 correspond à une gaussienne (bi-dimensionnelle) de moyenne $\boldsymbol{m}_1$ et de variance σ_1^2 (chaque composante de y est indépendante et de même variance σ_1^2).

En appliquant le test LRT (avec $\mu = 0$), trouve la règle de décision entre les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$. Montrer que ce test s'écrit sous la forme d'un réseau de neurones à une couche avec l'échelon d'Heavyside comme a fonction d'activation quand $\sigma_0^2 = \sigma_1^2$. Est-ce encore vrai si $\sigma_0^2 \neq \sigma_1^2$?

Exercice B.2. On considère le même modèle que l'exemple B.3 mais maintenant le signal x_n appartient à une M-PAM (cf. cours de COM105) d'amplitude A. Ecrire le test de Neyman-Pearson correspondant.

C. Solutions de certains exercices

Avertissement : la rédaction des solutions provient en partie d'un travail collaboratif avec des étudiants du cours MDI104 de Télécom Paris. Certaines erreurs sont susceptibles de subsister.

Exercice 2.6 1. N est à valeur dans $\mathbb{N}$, et F également.

De plus, le nombre de filles est inférieur au nombre d'enfants, donc $N \leq F$ presque sûrement.

Soit $(n, m) \in \mathbb{N}^2$, $m \leq n$,

$\mathbb{P}(N = n, F = m) = \mathbb{P}(N = n) \times \mathbb{P}(F = m | N = n)$

$\mathbb{P}(N = n) = e^{-\lambda} \frac{\lambda^n}{n!}$

De plus la probabilité d'avoir m filles parmi n est la même que celle de tirer m fois face dans n lancers de pièce, car les experiences sont iid.

Ainsi $\mathbb{P}(F = m | N = n) = \binom{n}{m}(\frac{1}{2})^n$

De plus, $\binom{n}{m} = \frac{n!}{m!(n-m)!}$

D'où $\forall (n, m) \in \mathbb{N}^2$, si $n \leq m, \mathbb{P}(N = n, F = m) = 0$,
sinon $\mathbb{P}(N = n, F = m) = e^{-\lambda} (\frac{\lambda}{2})^n \frac{1}{m!(n-m)!}$

2. Soit $m \in \mathbb{N}$.
$(N = n)_{n \in \mathbb{N}}$ est un système complet d'évenements, donc
$\mathbb{P}(F = m) = \sum_{n=0}^{\infty} \mathbb{P}(F = m, N = n)$

Or s'il y a m filles, il ne peut y avoir moins d'enfants.
Donc $\mathbb{P}(F = m) = \sum_{n=m}^{\infty} \mathbb{P}(F = m, N = n) = \frac{e^{-\lambda}}{m!} \sum_{n=m}^{\infty} (\frac{\lambda}{2})^n \frac{1}{(n-m)!} = \frac{e^{-\lambda}}{m!} (\frac{\lambda}{2})^m \sum_{n=0}^{\infty} (\frac{\lambda}{2})^n \frac{1}{n!}$
$= \frac{e^{-\lambda}}{m!} (\frac{\lambda}{2})^m e^{\lambda/2} = \frac{e^{-\lambda/2}}{m!} (\frac{\lambda}{2})^m$

M suit la même loi par symétrie du problème : $\mathbb{P}(M = m) = \frac{e^{-\lambda/2}}{m!} (\frac{\lambda}{2})^m$

3. On a simplement $M + F = N$.
Soit $(m, p) \in \mathbb{N}^2$.
$\mathbb{P}(F = m, M = p) = \mathbb{P}(F + M = m + p, M = p) = \mathbb{P}(N = m + p, M = p)$

Donc d'après q.1, $\mathbb{P}(F=m, M=p) = e^{-\lambda}\ (\frac{\lambda}{2})^{m+p}\frac{1}{m!p!}$

4. D'après les calculs précédents, $\forall (m,p) \in \mathbb{N}^2$,
$\mathbb{P}(F=m, M=p) = \mathbb{P}(F=m)\mathbb{P}(M=p)$.

   Ainsi le nombre de femelles pondues par la tortue est indépendant du nombre de males.

Exercice 2.7 On a les égalités suivantes :

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(\exp(X) \leq y)\\ &= \mathbb{P}(X \leq \log y)\\ &= F_X(\log y).\end{aligned}$$

En dérivant par rapport à $y \in \mathbb{R}_+^*$, on trouve finalement :

$$\forall y \in \mathbb{R}d+^*,\ f_Y(y) = \frac{1}{y} f_X(\log y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\log y - \mu)^2}{2\sigma^2}}.$$

Exercice 2.8 Non, aucun raison : ne pas confondre espérance et médiane. 50% des composants ont par définition une durée de vie inférieur à la médiane.

Exercice 2.11 1. Le charcutier ne peut pas vendre plus de choucroute qu'il n'en a fabriqué, donc le profit n'est **pas** $P = pX - cx$! On trouve la bonne formule en supposant qu'il vend exactement ce qu'il a produit puis en corrigeant selon la demande qu'il a reçue durant le journée :

$$P = (p-c)x - p(x-X)_+,$$

   en notant $x_+ = \max(0,x) = 0 \vee x$ la partie positive de x. En effet si le nombre de clients est supérieur à la quantité de choucroute prévue x, le charcutier vend tout son stock de la journée et réalise un chiffre d'affaire de $(p-c)x$ (on retire au prix le coût de production pour donner le bénéfice net). Si en revanche la demande est insuffisante pour écouler la quantité de choucroute préparée, alors on corrige le profit maximal possible ($(p-c)x$) avec la différence invendue, d'où le terme $-p(x-X)_+$. On remarque que si la demande est plus forte que prévue *i.e.* $x \leq X$, alors on retrouve $P = pX - cx$.

2. Par linéarité de l'espérance, les quantités en jeu étant supposées exister (il suffit que X soit intégrable pour $(X-x)_+$ le soit aussi par exemple), on a :

$$\mathbb{E}[P] = (p-c)x - p\mathbb{E}\big[(X-x)_+\big].$$

   On voit aisément que presque sûrement on a :

$$\begin{aligned}\int_0^x \mathbf{1}_{\{X\leq t\}}\mathrm{d}t &= \big[\int_0^x \mathbf{1}_{\{X\leq t\}}\mathrm{d}t\big]\mathbf{1}_{\{X\leq x\}}\\ &= \big[\int_0^x \mathbf{1}_{\{X\leq t\}}\mathrm{d}t\big]\mathbf{1}_{\{X\leq x\}}\\ &= \big[\int_{t\geq X} 1\mathrm{d}t\big]\mathbf{1}_{\{X\leq x\}}\\ &= (x-X)\mathbf{1}_{\{X\leq x\}}\\ &= (x-X)_+.\end{aligned}$$

Par Fubini-Tonelli (la fonction $(\omega,t)\mapsto \mathbf{1}_{\{X(\omega)\leq t\}}$ est mesurable et positive presque sûrement), on peut alors écrire [1] :

$$\begin{aligned}\mathbb{E}\big[\int_0^x \mathbf{1}_{\{X\leq t\}}\mathrm{d}t\big] &= \int_0^x \mathbb{E}[\mathbf{1}_{\{X\leq t\}}]\mathrm{d}t\\ &= \int_0^x \mathbb{P}(X\leq t)\mathrm{d}t\\ &= \int_0^x F_X(t)\mathrm{d}t.\end{aligned}$$

En recollant les morceaux, on trouve :

$$\mathbb{E}[P] = (p-c)x - p\int_0^x F_X(t)\mathrm{d}t.$$

Donc $\alpha = p-c$ et $\beta = p$.

3. On cherche les points critiques de $\varphi : x\mapsto \varphi(x) = \mathbb{E}[P]$ sur $\mathbb{R}_+$ (la quantité de choucroute produite x est positive !). Cette fonction est dérivable car F_X est supposée continue, et donc son intégrale de 0 à x est $\mathcal{C}^1$. On étudie le sens de variation de φ :

$$\varphi'(x)\geq 0 \Leftrightarrow (p-c) - pF_X(x)\geq 0,$$

ou encore $x\geq F^{-1}(\frac{p-c}{p})$ en supposant que l'équation $F(x) = (p-c)/p$ admet une unique solution). Remarquons que $(p-c)/p\in]0,1[$ puisque $c\in]0,p[$, donc sans hypothèse autre que le continuité de F_X, on a toujours une solution au moins à cette équation. Finalement la quantité à produire qui maximise le chiffre d'affaire moyen est $x = F^{-1}(\frac{p-c}{p})$.

4. Il s'agit de trouver a tel que :

$$\mathbb{P}(P\geq a) = \mathbb{P}\big((x-X)_+\leq (p-c)x - a\big) = 0.95.$$

Dit autrement, il s'agit d'exprimer $F_{(x-X)_+}$ en fonction de F_X (et de c et de p). Par un conditionnement qui doit être évident maintenant, pour $t\geq 0$:

1. Rappel : pour un événement A quelconque, $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$.

$$\begin{aligned} F_{(x-X)_+}(t) &= \mathbb{P}\big((x-X)_+ \leq t\big) \\ &= \mathbb{P}\big((x-X)_+ \leq t, X \leq x\big) + \mathbb{P}\big((x-X)_+ \leq t, X > x\big) \\ &= \mathbb{P}(x - X \leq t, X \leq x) + \mathbb{P}\big(0 \leq t, X > x\big) \\ &= \mathbb{P}(x - t \leq X \leq x) + \mathbb{P}\big(X > x\big) \\ &= F_X(x) - F_X(x-t) + (1 - F_X(x)) \\ &= 1 - F_X(x-t). \end{aligned}$$

On remarque que si $x \leq t$, alors cette dernière quantité vaut 1 (X est à valeurs positives). C'est cohérent : dans ce cas $(x-X)_+ \leq (t-X)_+ \leq t$. Finalement :

$$\mathbb{P}(P \geq a) = 1 - F_X\big((p-c)x - a - t\big),$$

et la valeur a recherchée est $a = (p-c)x - t - F_X^{-1}(0.05)$. Cela s'interprète comme le profit minimal réalisé avec probabilité 0.95 **une fois que la quantité de choucroute à produire x est déterminée**.

5. Par ce qui précède, on a

$$x = -\frac{1}{\lambda}\log\big(\frac{p-c}{p}\big),$$

et :

$$a = -\frac{p-c}{\lambda}\log\big(\frac{p-c}{p}\big) - t + \frac{1}{\lambda}\log(0.95).$$

Exercice 2.12 On donne la réponse à la deuxième question. On a

$$\begin{aligned} \mathbb{E}(Z) &= \mathbb{E}(YB) + \mathbb{E}(X(1-B)) \\ &= \mathbb{E}(Y)\mathbb{E}(B) + \mathbb{E}(X)(1 - \mathbb{E}(B)) \text{ par independance} \end{aligned}$$

or on sait que pour une v.a X qui suit une loi exponentielle de parametre λ , $\mathbb{E}(X) = \dfrac{1}{\lambda}$ donc

$$\begin{aligned} \mathbb{E}(Z) &= \frac{2}{\lambda}\frac{3}{4} + \frac{1}{\lambda}(1 - \frac{3}{4}) \\ &= \frac{7}{4\lambda} \geq \frac{1}{\lambda} = \mathbb{E}(X) \end{aligned}$$

Donc le traitement allonge bien l'esperance de vie des patients

Exercice 2.13 Par le théorème de transfert, $\mathbb{P}(X^2+Y^2 \leq 1) = \int\int \mathbf{1}_D(x,y)p_{X,Y}(x,y)dxdy$, où D représente le disque unité, et où $p_{X,Y}(x,y)$ est la densité jointe du couple X, Y. Par l'énoncé, cette loi est l'indicatrice du carré unité $C = [O,1] \times [0,1]$. Finalement, $\mathbb{P}(X^2+Y^2 \leq 1) = \int\int \mathbf{1}_D(x,y)\mathbf{1}_C(x,y)dxdy$. Donc, $\mathbb{P}(X^2+Y^2 \leq 1) = \int\int \mathbf{1}_{D\cap C}(x,y)dxdy$. Ainsi, $\mathbb{P}(X^2+Y^2 \leq 1)$ est simplement l'aire de $D \cap C$, qui vaut $\pi/4$.

Exercice 2.14 1. Par definition d'une probabilite , on sait que :

$$1 = \mathbb{P}((X,Y) \in \mathbb{R}^2) = \iint \alpha 1_C(x,y)dxdy = \alpha\lambda_2(C) = \alpha\pi$$

d'ou $\alpha = \dfrac{1}{\pi}$

2. On a

$$f_X(x) = \int f_{(X,Y)}(x,y)dy = \int \frac{1}{\pi}1_C(x,y)dy \overset{*}{=} \frac{1}{\pi}\int 1_{[-\sqrt{1-x^2},\sqrt{1-x^2}]}(y)1_{[-1,1]}(x)dy$$

enfin on obtient $f_X(x) = \dfrac{2}{\pi}\sqrt{1-x^2}1_{[-1,1]}(x)$
Montrons $*$, en effet

$$(x,y) \in C \Leftrightarrow x^2+y^2 \leq 1 \Leftrightarrow |y| \leq \sqrt{1-x^2} \Leftrightarrow y \in [-\sqrt{1-x^2}, \sqrt{1-x^2}], x \in [-1,1]$$

x et y jouent un role symetrique , on en deduit que $f_Y(y) = \dfrac{2}{\pi}\sqrt{1-y^2}1_{[-1,1]}(y)$

3. On remarque que $f_{(X,Y)}(x,y) \neq f_X(x)f_Y(y)$, donc X et Y ne sont pas independantes.

Exercice 2.15 1. on a $T = 2\pi - |V-U|$

2. Calculons $\mathbb{E}(|U-V|)$
Grâce à une astuce qu'on retrouve souvent en probabilité, on écrit

$$\begin{aligned}
\mathbb{E}(|U-V|) &= \mathbb{E}(\mathbf{1}_{U\geq V}.(U-V)) + \mathbb{E}(\mathbf{1}_{U<V}(V-U)) \\
&= 2\mathbb{E}(\mathbf{1}_{U>V}(U-V)) \text{ par symetrie de } U \text{ et } V \\
&= 2\mathbb{E}(\mathbf{1}_\Delta(U,V).(U-V)) \text{ avec } \Delta = (u,v), u > v \\
&= \frac{2}{(2\pi)^2}\iint \mathbf{1}_\Delta(u,v).(u-v)\mathbf{1}_{[0,2\pi]}(u)\mathbf{1}_{[0,2\pi]}(v)dudv \\
&= \frac{1}{2\pi^2}\int_0^{2\pi}\int_0^u (u-v)dvdu \\
&= \frac{2\pi}{3}
\end{aligned}$$

puis enfin

$$\mathbb{E}(T) = 2\pi - \mathbb{E}(|U-V|) = \frac{4\pi}{3}$$

ce resultat est *a priori* surprenant dans la mesure où la réponse à laquelle on s'attend est π

Exercice 4.5 1. On a $\mathbb{P}_\theta(T=k) = (1-\theta)^{k-1}\theta$, c'est une loi géométrique de paramètre θ sur $\mathbb{N}^*$.

2. La vraisemblance est donnée par :

$$L_\theta(T_1,\ldots,T_n) = \prod_{i=1}^n (1-\theta)^{T_i-1}\theta = (1-\theta)^{\sum_{i=1}^n T_i - n}\theta^n\,.$$

Après passage au logarithme, on résoud :

$$\begin{aligned} 0 &= \frac{d}{d\theta} \ln L_\theta(T_1, \ldots, T_n) \\ &= \frac{d}{d\theta} \left[(\sum_{i=1}^{n} T_i - n) \ln(1-\theta) + n \ln \theta \right] \\ &= -(\sum_{i=1}^{n} T_i - n) \frac{1}{1-\theta} + \frac{n}{\theta} \,. \end{aligned}$$

On multiplie par $\theta(1-\theta)$:

$$\begin{aligned} 0 &= -(\sum_{i=1}^{n} T_i - n)\theta + n(1-\theta) \\ &= -\sum_{i=1}^{n} T_i \theta + n \,, \end{aligned}$$

et l'estimateur du maximum de vraisemblance est donc donné par :

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} T_i} \,.$$

3. L'estimateur de l'espérance de survie est $1/\hat{\theta}$, soit $\sum_i T_i/n$. On retrouve l'estimateur de la moyenne empirique.
4. On a $\mathbb{P}_\theta(T \geq 24) = \sum_{k=24}^{\infty} (1-\theta)^{k-1}\theta = (1-\theta)^{23}$. L'estimateur est donné par $(1-\hat{\theta})^{23}$, où $\hat{\theta}$ est l'estimateur du maximum de vraisemblance.
5. To do.
6. L'espérance est nulle et la variance est égale à un, lorsque $\theta = 1/D$.