Statistique Mathématique
ENSAE - Mastères Spécialisés

Guillermo Durand

2024-2025

Table des matières

Chapitre 1

Probabilités

1.1 Théorie de la mesure

On commence par des rappels de théorie de la mesure, des notes succintes de l'essentiel à savoir sont disponibles ici : https://durandg12.github.io/files/notes_mesure.pdf. Comme ces notes ont été écrites pour un autre cours plus réduit où il n'y a pas de rappels de probabilités, certains éléments de ces notes sont répétés et étendus dans les sections qui suivent.

1.2 Espace Probabilisé

Définition 1. *Un espace probabilisé est un triplet $(\Omega, \mathcal{A}, \mathbb{P})$ où :*

- Ω *est un ensemble ;*
- $\mathcal{A}$ *une tribu sur* Ω *;*
- $\mathbb{P}$ *une mesure de probabilité sur* $\mathcal{A}$ *:*

$$\mathbb{P}(\varnothing) = 0, \quad \mathbb{P}(\Omega) = 1,$$
$$\forall \left\{ (A_i) \in \mathcal{A}^{\mathbb{N}} : \forall i \neq j, A_i \cap A_j = \varnothing \right\}, \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Un élément A de $\mathcal{A}$ est appelé un événement.
On peut démontrer les règles de calcul suivantes, qui sont bien utiles.

Proposition 1. *Soit $A, B \in \mathcal{A}$, soit $(A_i) \in \mathcal{A}^{\mathbb{N}}$ une suite d'événements. Alors :*

1. $\mathbb{P}(A^{\mathsf{c}}) = 1 - \mathbb{P}(A)$*, où A^{c} désigne le complémentaire de A dans Ω, parfois aussi noté $\bar{A}$ (mais à ne pas confondre avec l'adhérence),*
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$*,*
3. $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$*,*
4. *si $A \subseteq B$, $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$ et en particulier $\mathbb{P}(A) \leq \mathbb{P}(B)$,*
5. *si de plus la suite est croissante : $A_i \subseteq A_{i+1}$ pour tout $i \in \mathbb{N}$, $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_i)$,*
6. *si de plus la suite est décroissante : $A_{i+1} \subseteq A_i$ pour tout $i \in \mathbb{N}$, $\mathbb{P}\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} \mathbb{P}(A_i)$.*

1.3 Variables aléatoires réelles

Définition 2. *X est une variable aléatoire si X est une fonction mesurable de $(\Omega, \mathcal{A})$ dans un ensemble mesuré $(\mathbb{X}, \mathfrak{F})$.*

Définition 3. *La loi d'une variable aléatoire X, notée $\mathcal{L}(X)$, $\mathbb{P}_X$ ou encore $X_{\#\mathbb{P}}$ est la mesure image de $\mathbb{P}$ par X, qu'on appelle aussi mesure induite : $\forall F \in \mathfrak{F}, \mathbb{P}_X(F) = \mathbb{P}\left(X^{-1}(F)\right) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in F\}) = \mathbb{P}(X \in F)$. On dit encore que X suit $\mathcal{L}(X)$ et on note $X \sim \mathcal{L}(X)$.*

Définition 4. *X est une variable aléatoire réelle si $(\mathbb{X}, \mathfrak{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*

La fonction de répartition (en anglais cumulative distribution function ou cdf) caractérise entièrement la loi d'une variable aléatoire réelle. Elle est la fonction :

$$\begin{aligned} F_X : \mathbb{R} &\longrightarrow [0,1] \\ x &\longmapsto \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X^{-1}(]-\infty, x])) = \mathbb{P}(X \leq x) \end{aligned}$$

La fonction de survie, ou fonction de queue de distribution, est la fonction $x \mapsto 1 - F_X(x) = \mathbb{P}(X > x)$. Elle caractérise aussi la loi.

Exemple : si X est la fonction constante égale à $e \in \mathbb{R}$, sa loi est la mesure de Dirac δ_e, c'est la mesure de définie par $\delta_e(B) = \mathbb{1}_{e \in B}$. En effet, $\mathbb{P}_X(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) = \mathbb{P}(\{\omega : e \in B\}) = \mathbb{P}(\Omega) = 1$ si $e \in B$ et $= \mathbb{P}(\varnothing) = 0$ sinon. Sa cdf est donnée par $F_X(x) = \mathbb{1}_{e \in]-\infty, x]} = \mathbb{1}_{e \geq x}$.

Exemple : on lance deux fois d'affilée, indépendamment, une pièce équilibrée. $\Omega = \{PP, FP, PF, FF\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, et $\mathbb{P}$ est la probabilité uniforme sur Ω : $\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{|A|}{4}$ pour tout $A \in \mathcal{A}$. Soit X la variable aléatoire qui compte le nombre de faces : $X(PP) = 0$, $X(FP) = X(PF) = 1$, $X(FF) = 2$. Sa loi est donnée par

$$\begin{aligned} \mathbb{P}_X(B) &= \mathbb{P}(X \in B) \\ &= \mathbb{P}(X \in (B \cap \{0\}) \cup (B \cap \{1\}) \cup (B \cap \{2\})) \\ &= \mathbb{P}(X \in B \cap \{0\}) + \mathbb{P}(X \in B \cap \{1\}) + \mathbb{P}(X \in B \cap \{2\}) \\ &= \mathbb{P}(X = 0)\mathbb{1}_{0 \in B} + \mathbb{P}(X = 1)\mathbb{1}_{1 \in B} + \mathbb{P}(X = 2)\mathbb{1}_{2 \in B} \\ &= \mathbb{P}(\{PP\})\mathbb{1}_{0 \in B} + \mathbb{P}(\{FP, PF\})\mathbb{1}_{1 \in B} + \mathbb{P}(\{FF\})\mathbb{1}_{2 \in B} \\ &= \frac{1}{4}\mathbb{1}_{0 \in B} + \frac{2}{4}\mathbb{1}_{1 \in B} + \frac{1}{4}\mathbb{1}_{2 \in B} \\ &= \frac{1}{4}\mathbb{1}_{0 \in B} + \frac{1}{2}\mathbb{1}_{1 \in B} + \frac{1}{4}\mathbb{1}_{2 \in B}. \end{aligned}$$

C'est la loi binomiale de paramètres 2 et $\frac{1}{2}$ notée $\mathcal{B}\left(2, \frac{1}{2}\right)$. Sa fonction de répartition est donnée par $F_X(x) = \frac{1}{4}\mathbb{1}_{0 \leq x} + \frac{1}{2}\mathbb{1}_{1 \leq x} + \frac{1}{4}\mathbb{1}_{2 \leq x}$.

Proposition 2. *F_X est une fonction croissante, continue à droite et limitée à gauche (càdlàg), qui a pour limite 0 en $-\infty$ et 1 en $+\infty$.*

Démonstration. Soit $x \leq y$, $]-\infty, x] \subseteq]-\infty, y]$ et donc $F_X(x) = \mathbb{P}(X^{-1}(]-\infty, x])) \leq \mathbb{P}(X^{-1}(]-\infty, x])) = F_X(y)$. Par croissance F_X est limitée à gauche et à droite en tout point. De plus, pour $x \in \mathbb{R}$, pour toute suite (ε_n) telle que $\varepsilon_n > 0$ et $\varepsilon_n \underset{n\to\infty}{\longrightarrow} 0$, $]-\infty, x] = \bigcap_{n=1}^{\infty}]-\infty, x + \varepsilon_n]$ et l'union est décroissante donc $\mathbb{P}(X^{-1}(]-\infty, x])) = \lim_{n\to\infty} \mathbb{P}(X^{-1}(]-\infty, x + \varepsilon_n]))$ soit $F_X(x) = \lim_{n\to\infty} F_X(x + \varepsilon_n)$ ce qui est exactement la continuité à droite de F_X. Soit (A_n) strictement croissante qui tend vers ∞ et (B_n) strictement décroissante qui tend vers $-\infty$. On a $\mathbb{R} = \bigcup_{n=1}^{\infty}]-\infty, A_n]$ avec l'union qui est croissante et $\varnothing = \bigcap_{n=1}^{\infty}]-\infty, B_n]$ avec l'union qui est décroissante et donc par les mêmes arguments on obtient les limites en $\pm\infty$. □

Réciproquement, toute fonction avec ces quatre propriétés définit une mesure de probabilité sur $\mathbb{R}$ et est donc la cdf d'une certaine variable aléatoire réelle. On note F_X^- sa fonction limite à gauche associée : $F_X^-(x) = \lim_{\substack{\varepsilon\to 0 \\ \varepsilon > 0}} F_X(x - \varepsilon)$.

Vu que la cdf caractérise la loi d'une variable aléatoire, pour montrer que X suit $\mathcal{L}$, on peut calculer sa cdf et montrer qu'elle est égale à la cdf d'une variable aléatoire suivant $\mathcal{L}$.

Exemple : montrons que la loi de aX, $a > 0$, $X \sim \mathcal{E}(\lambda)$, est la loi $\mathcal{E}\left(\frac{\lambda}{a}\right)$. On a $\mathbb{P}(X \leq x) = (1 - \exp(-\lambda x))\mathbb{1}_{x>0}$, donc

$$\begin{aligned} \mathbb{P}(aX \leq x) &= \mathbb{P}\left(X \leq \frac{x}{a}\right) \\ &= \left(1 - \exp\left(-\frac{\lambda}{a}x\right)\right)\mathbb{1}_{x>0}, \end{aligned}$$

on reconnaît bien la cdf de $\mathcal{E}\left(\frac{\lambda}{a}\right)$.

Variables aléatoires discrètes

Une variable aléatoire discrète prend ses valeurs dans un ensemble au plus dénombrable (donc, paradoxalement, pas forcément discret), elles sont notées ici $(x_i)_{i\in I}$, $I \subset \mathbb{N}$. Sa loi est $\mathbb{P}_X = \sum_{i\in I} p_i \delta_{x_i}$, où les p_i sont des réels positifs tels que $\sum_{i=1} p_i = 1$. En particulier $p_i = \mathbb{P}(X = x_i)$.

Sa densité par rapport à la mesure de comptage sur l'ensemble $\{x_i : i \in I\}$, notée $\sum_{i\in I} \delta_{x_i}$ (qui est bien σ-finie car I est au plus dénombrable), est aussi appelée fonction de masse de probabilité (en anglais probability mass function ou pmf) et correspond aux probabilités des atomes : $\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\sum_{i\in I}\delta_{x_i}}(x_i) = \mathbb{P}(X = x_i) = p_i$, et $\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\sum_{i\in I}\delta_{x_i}}(x) = 0$ partout ailleurs. On a alors :

$$\forall x \in \mathbb{R},\ F_X(x) = \sum_{i\in I} p_i \mathbb{1}_{x_i \le x},$$

et réciproquement

$$\forall i \in I,\ p_i = F_X(x_i) - F_X^-(x_i).$$

En particulier la pmf caractérise la loi.

Variables aléatoires absolument continues

Une variable aléatoire (absolument, on omettra désormais ce mot) continue est une loi dont la mesure induite est dominée par la mesure de Lebesgue, donc elle admet une densité définie sur $\mathbb{R}$, parfois notée $f_X = \frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\lambda}$ par rapport à la mesure de Lebesgue λ. On a alors :

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t$$

et réciproquement on peut définir une densité f_X par $f_X(x) = F_X'(x)$ pour tout x où F_X est dérivable (et en prenant une valeur quelconque ailleurs).

En particulier la densité caractérise la loi. En particulier aussi, F_X est une fonction continue. Attention, la réciproque n'est pas vraie : prendre par exemple la fonction de Cantor qui est continue mais non-dérivable sur l'ensemble de Cantor qui est indénombrable et de mesure de Lebesgue nulle. Son support (cf § suivant) est justement l'ensemble de Cantor donc la mesure associée ne peut pas être dominée par la mesure de Lebesgue.

Support

Le support d'une variable aléatoire réelle est, informellement, l'ensemble des valeurs qu'elle peut prendre.

Formellement, il est défini de manière générale comme

$$\begin{aligned}\mathrm{supp}(X) &= \{x \in \mathbb{R} : \forall O \text{ ouvert tel que } x \in O, \mathbb{P}_X(O) > 0\}\\ &= \{x \in \mathbb{R} : \forall \varepsilon > 0, \mathbb{P}_X(]x-\varepsilon, x+\varepsilon[) > 0\}\\ &= \{x \in \mathbb{R} : \forall \varepsilon > 0, \mathbb{P}(x-\varepsilon < X < x+\varepsilon) > 0\}.\end{aligned}$$

On peut montrer avec des propriétés sur la topologie de $\mathbb{R}$ que c'est aussi le plus petit (au sens de l'inclusion) fermé C tel que $\mathbb{P}_X(C) = 1$:

$$\mathrm{supp}(X) = \bigcap_{\substack{C \text{ fermé}\\ \mathbb{P}_X(C)=1}} C.$$

Si la loi de X est dominée par une mesure μ et possède donc une densité f_X par rapport à cette mesure μ, on peut aussi montrer que c'est le support essentiel de f_X défini par

$$\mathrm{ess\,supp}(f_X) = \left(\bigcup_{\substack{U \text{ ouvert}\\ f_X=0\ \mu\text{-p.p. sur } U}} U\right)^{\mathsf{c}}.$$

Pour une variable discrète, on peut montrer que c'est l'ensemble $\{x_i : i \in I \text{ et } p_i > 0\}$.

Pour les variables absolument continues, naturellement on veut considérer que c'est l'ensemble où sa densité est > 0 : $\{x : f_X(x) > 0\}$ ou l'adhérence ou l'intérieur de cet ensemble ou une combinaison de ces opérations mais ça ne marche qu'à condition de prendre une "bonne" densité. Exemple : le support de $\mathcal{U}([0,1])$ est $[0,1]$ et une densité de cette loi est $f_X(x) = \mathbb{1}_{(\mathbb{R}\setminus\mathbb{Q})\cap]0,1[}(x) + \mathbb{1}_{\mathbb{Q}\setminus[0,1]}(x)$ mais $\overline{\{x : f_X(x) > 0\}} = \mathbb{R}$ et $\{x : f_X\mathring{(}x) > 0\} = \varnothing$. Si on prend par contre comme densité $g_X(x) = \mathbb{1}_{[0,1]}(x)$ alors on a bien $[0,1] = \{x : g_X(x) > 0\}$. En fait en général, si on prend la densité définie par $h_X(x) = F'_X(x)$ là où F_X est dérivable, et 0 ailleurs, alors on peut montrer que $\text{supp}(X) = \overline{\{x : h_X(x) > 0\}}$ (dans l'exemple uniforme $h_X(x) = \mathbb{1}_{]0,1[}(x)$).

1.4 Moments : espérance, variance, etc.

L'espérance (ou moyenne), si elle existe, est la quantité :

$$\mathbb{E}[X] = \int X(\omega)\mathrm{d}\mathbb{P}(\omega) = \int x\mathrm{d}\mathbb{P}_X(x) = \begin{cases} \sum_{i\in I} x_i \mathbb{P}(X = x_i) & \text{si X est discrète,} \\ \int_{-\infty}^{+\infty} x f_X(x)\mathrm{d}x & \text{si X est continue.} \end{cases}$$

Plus généralement, pour toute fonction φ mesurable, on a, si elle existe :

$$\mathbb{E}[\varphi(X)] = \int \varphi(x)\mathrm{d}\mathbb{P}_X(x) = \begin{cases} \sum_{i\in I} \varphi(x_i) \mathbb{P}(X = x_i) & \text{si X est discrète,} \\ \int_{-\infty}^{+\infty} \varphi(x) f_X(x)\mathrm{d}x & \text{si X est continue.} \end{cases}$$

L'espérance de X existe si et seulement si $\mathbb{E}[|X|] < \infty$, si et seulement si X est une fonction intégrable par rapport à la mesure $\mathbb{P}$ sur Ω. Dans ce cas on dit que $X \in L^1(\Omega)$ ou plus simplement $X \in L^1$.

L'espérance est linéaire : $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ si $a, b \in \mathbb{R}$, $X, Y \in L^1$.

Si $\mathbb{E}\left[|X|^k\right] < \infty$, on dit que X possède un moment d'ordre k, noté $m^{(k)} = \mathbb{E}\left[X^k\right]$ et que $X \in L^k(\Omega)$ ou plus simplement $X \in L^k$.

La variance est le moment centré d'ordre 2, i.e.

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = m^{(2)} - \left(m^{(1)}\right)^2$$

La variance est 2-homogène : $\mathbb{V}(aX) = a^2\mathbb{V}(X)$ si $X \in L^2$. La variance est additive sous indépendance : $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)$ si $X, Y \in L^2$ et si X et Y sont indépendantes (à montrer en exercice). L'écart-type est la racine carrée de la variance, on les note souvent respectivement σ et σ^2. Ainsi $\sigma = \sqrt{\mathbb{V}(X)}$.

Plus généralement, si $X \in L^k$, son moment centré d'ordre k est défini par $\mathbb{E}\left[(X - \mathbb{E}[X])^k\right]$.

Une méthode pour trouver la densité de $\psi(X)$ quand X est une variable continue de densité f_X connue est la méthode de la fonction muette. Il faut que ψ soit un C^1-difféomorphisme.

Proposition 3. *Soit Y une v.a. continue. Si, pour tout h continue bornée, $\mathbb{E}[h(Y)] = \int h(y)g(y)\mathrm{d}y$, alors g est une densité de Y.*

Donc pour trouver une densité de $\psi(X)$, on prend h continue bornée, on calcule $\mathbb{E}[h(\psi(X))] = \int h(\psi(x))f_X(x)\mathrm{d}x$, et on applique la formule de changement de variable

$$\mathbb{E}[h(\psi(X))] = \int h(y) f_X(\psi^{-1}(y)) \left|\det J_{\psi^{-1}}(y)\right| \mathrm{d}y$$

avec $J_{\psi^{-1}}$ la Jacobienne de ψ^{-1} (juste sa dérivée en une dimension). Ainsi une densité de $\psi(X)$ est $y \mapsto f_X(\psi^{-1}(y)) \left|\det J_{\psi^{-1}}(y)\right|$.

Exemple : montrons (encore) que la loi de aX, $a > 0$, $X \sim \mathcal{E}(\lambda)$, est la loi $\mathcal{E}\left(\frac{\lambda}{a}\right)$. On a une densité de $\mathcal{E}(\lambda)$ qui est $f_X : x \mapsto \lambda \exp(-\lambda x)\mathbb{1}_{x>0}$. Soit h continue bornée.

$$\begin{aligned} \mathbb{E}[h(aX)] &= \int h(ax) f_X(x)\mathrm{d}x \\ &= \int_0^\infty h(\psi(x))\lambda \exp(-\lambda x)\mathrm{d}x \end{aligned}$$

avec $\psi(x) = ax$, inversible d'inverse $\psi^{-1} : y \mapsto \frac{y}{a}$ de Jacobienne $J_{\psi^{-1}}(y) = \frac{1}{a}$. Donc

$$\begin{aligned}\mathbb{E}\left[h(aX)\right] &= \int_0^\infty h(\psi(x))\lambda \exp(-\lambda x)\mathrm{d}x \\ &= \int_0^\infty h(y)\frac{\lambda}{a}\exp\left(-\frac{\lambda}{a}y\right)\mathrm{d}y \\ &= \int h(y)\frac{\lambda}{a}\exp\left(-\frac{\lambda}{a}y\right)\mathbb{1}_{y>0}\mathrm{d}y.\end{aligned}$$

Donc une densité de aX est donnée par $y \mapsto \frac{\lambda}{a}\exp(-\frac{\lambda}{a}y)\mathbb{1}_{y>0}$, on reconnaît une densité de la loi $\mathcal{E}\left(\frac{\lambda}{a}\right)$. Informellement, pour le changement de variable en général on n'introduit pas ψ et on rédige plutôt de la façon suivante : "on pose $y = ax$, $\frac{\mathrm{d}y}{\mathrm{d}x} = a$ donc $\mathrm{d}x = \frac{1}{a}\mathrm{d}y$..."

1.5 Couples et vecteurs aléatoires

Un couple aléatoire (X, Y) est une variable aléatoire à valeur dans $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. La covariance entre X et Y est

$$\begin{aligned}\mathbb{C}\mathrm{ov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \text{ (à démontrer).}\end{aligned}$$

La covariance est symétrique.

Un vecteur aléatoire est une variable aléatoire à valeur dans $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. Un vecteur peut être discret ou continu (ou ni l'un ni l'autre, mais on ne traitera pas ce cas ici) et dans ces cas, leur densité est analogue au cas réel. La notion de support est aussi analogue au cas réel.

L'espérance est un vecteur ayant le nombre de composantes correspondantes :

$$\mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^\top.$$

La variance est alors une matrice (dite de variance-covariance), symétrique semi-définie positive, qui se définit par :

$$\begin{aligned}\Sigma_X &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top] \\ &= \begin{pmatrix} \mathbb{V}(X_1) & \mathbb{C}\mathrm{ov}(X_1, X_2) & \ldots & \mathbb{C}\mathrm{ov}(X_1, X_n) \\ \mathbb{C}\mathrm{ov}(X_1, X_2) & \mathbb{V}(X_2) & \ldots & \mathbb{C}\mathrm{ov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\mathrm{ov}(X_n, X_1) & \mathbb{C}\mathrm{ov}(X_n, X_2) & \ldots & \mathbb{V}(X_n) \end{pmatrix}.\end{aligned}$$

Proposition 4. *Soit X vecteur aléatoire de taille n, $A \in \mathcal{M}_{p,n}(\mathbb{R})$ et $b \in \mathbb{R}^p$, alors $\mathbb{E}\left[AX + b\right] = A\mathbb{E}\left[X\right] + b$ et $\Sigma_{AX+b} = A\Sigma_X A^\top$.*

Démonstration. Pour l'espérance c'est juste la linéarité de l'espérance, à écrire coordonnée par coordonnée.

$$\begin{aligned}\Sigma_{AX+b} &= \mathbb{E}[(AX + b - \mathbb{E}[AX + b])(AX + b - \mathbb{E}[AX + b])^\top] \\ &= \mathbb{E}[(AX - \mathbb{E}[AX])(AX - \mathbb{E}[AX])^\top] \\ &= \mathbb{E}[A(X - \mathbb{E}[X])(A(X - \mathbb{E}[X]))^\top] \\ &= \mathbb{E}[A(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top A^\top] \\ &= A\Sigma_X A^\top.\end{aligned}$$

□

Indépendance

Un vecteur $X = (X_1, \dots, X_n)$ est à composantes indépendantes si et seulement si :

$$\forall (B_1, \dots, B_n) \in \mathcal{B}(\mathbb{R})^n,\ \mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mathbb{P}(X_1 \in B_1) \times \cdots \times \mathbb{P}(X_n \in B_n).$$

Attention ! Ce n'est pas équivalent à l'indépendance deux à deux des X_i.

Si X et Y sont indépendantes, $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ et donc $\mathbb{C}\mathrm{ov}(X, Y) = 0$. Attention ! La réciproque est fausse !

En termes de mesure induite, cela veut dire que la mesure induite de X est la mesure produit des mesures induites des X_i, ce qu'on note $\mathbb{P}_X = \mathbb{P}_{X_1} \otimes \cdots \otimes \mathbb{P}_{X_n} = \bigotimes_{i=1}^n \mathbb{P}_{X_i}$.

Si X est à composantes indépendantes et $\mathbb{P}_{X_i} \ll \mu_i$ pour tout i, alors $\mathbb{P}_X \ll \bigotimes_{i=1}^n \mu_i$ et $\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\bigotimes_{i=1}^n \mu_i} : (x_1, \dots, x_n) \mapsto \prod_{i=1}^n \frac{\mathrm{d}\mathbb{P}_{X_i}}{\mathrm{d}\mu_i}(x_i)$ est une densité de X. Attention, sans indépendance on n'a même pas forcément que $\mathbb{P}_X \ll \bigotimes_{i=1}^n \mu_i$: exemple si $X \sim \mathcal{N}(0,1)$, alors $\mathbb{P}_{(X,X)} \not\ll \mathrm{Leb}(\mathbb{R}^2)$ (en effet le support de (X, X) est la droite d'équation $y = x$ qui est de mesure de Lebesgue nulle dans $\mathbb{R}^2$).

Définition 5. *Un n-échantillon est un vecteur aléatoire $X = (X_1, \dots, X_n)$ à composantes indépendantes identiquement distribuées (iid), sa mesure induite est donc la mesure produit $\mathbb{P}_X = \mathbb{P}_{X_1}^{\otimes n}$.*

Vecteurs gaussiens

Un vecteur aléatoire $(X_1, \dots, X_n)$ d'espérance $m \in \mathbb{R}^n$ et de matrice de variance-covariance $\Sigma_X \in \mathcal{M}_{n,n}(\mathbb{R})$ est gaussien si toute combinaison linéaire de ses coordonnées est une v.a.r. gaussienne. Dans ce cas on note $(X_1, \dots, X_n) \sim \mathcal{N}(m, \Sigma_X)$. En particulier m et Σ_X caractérisent la loi de X.

Proposition 5. *Si $X \sim \mathcal{N}(m, \Sigma)$ est gaussien en dimension n et si $A \in \mathcal{M}_{p,n}(\mathbb{R})$ et $b \in \mathbb{R}^p$ alors $AX + b \sim \mathcal{N}(A\mu + b, A\Sigma_X A^\top)$.*

Proposition 6. *Si les X_i sont des gaussiennes et les composantes de $X = (X_1, \dots, X_n)$ sont indépendantes, alors X est gaussien.*

Proposition 7. *Si $X = (X_1, \dots, X_n)$ est gaussian, alors X_i et X_j sont indépendantes si et seulement si $\mathbb{C}\mathrm{ov}(X_i, X_j) = 0$.*

C'est une réciproque à la propriété ci-dessus qui n'est pas vraie en général mais qui l'est donc pour les vecteurs gaussiens.

Théorème 1 (Théorème de Cochran)**.** *Soit $X = (X_1, \dots, X_n) \sim \mathcal{N}(m, \sigma^2 \mathrm{Id}_n)$, $m \in \mathbb{R}^n$, $\sigma^2 > 0$. Soient $F_1, \dots, F_r$ des sous-espaces orthogonaux deux à deux de $\mathbb{R}^n$ dont la somme fait $\mathbb{R}^n$. On note, pour $1 \le i \le r$, P_{F_i} la matrice de la projection orthogonale sur F_i et d_i la dimension de F_i. Alors*

1. *$P_{F_1}X, \dots, P_{F_r}X$ sont indépendants deux à deux.*
2. *$\frac{\left\|P_{F_i}(X-m)\right\|^2}{\sigma^2} \sim \chi^2(d_i)$, la loi du χ^2 à d_i degrés de libertés.*

Le corollaire suivant du théorème est très utile :

Corollaire 1 (Corollaire du théorème de Cochran)**.** *Soit $X = (X_1, \dots, X_n)$ avec les X_i indépendantes de loi $\mathcal{N}(m, \sigma^2)$, $m \in \mathbb{R}$, $\sigma^2 > 0$. On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Alors $\overline{X}_n$ et $\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ sont indépendantes, et $\frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2} \sim \chi^2(n-1)$.*

Démonstration. Pour se remettre dans le contexte du théorème, on a donc ici $X = (X_1, \dots, X_n) \sim \mathcal{N}\left(m \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \sigma^2 \mathrm{Id}_n\right)$. On pose $r = 2$, F_1 l'espace engendré par le vecteur $\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, et F_2 sont orthogonal dans $\mathbb{R}^n$. Alors $\dim(F_2) = n - \dim(F_1) = n - 1$,

$$P_{F_1} = \frac{1}{n} \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix},$$

$P_{F_1}X = \overline{X}_n \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$, $P_{F_2} = \mathrm{Id}_n - P_{F_1}$, et $P_{F_2}X = X - \overline{X}_n \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \begin{pmatrix}X_1 - \overline{X}_n\\ \vdots \\ X_n - \overline{X}_n\end{pmatrix}$. L'indépendance de $\overline{X}_n$ et $\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$ se déduit donc du premier point du théorème. De plus, le deuxième point du théorème affirme que

$$\frac{\left\| P_{F_2}\left(X - m\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right)\right\|^2}{\sigma^2} \sim \chi^2(\dim(F_2)) = \chi^2(n-1),$$

or $m \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \in F_1$ donc $P_{F_2} m \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = 0$ et enfin

$$\begin{aligned}\frac{\left\| P_{F_2}\left(X - m\begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}\right)\right\|^2}{\sigma^2} &= \frac{\|P_{F_2}X\|^2}{\sigma^2}\\ &= \frac{\left\|\begin{pmatrix}X_1 - \overline{X}_n\\ \vdots \\ X_n - \overline{X}_n\end{pmatrix}\right\|^2}{\sigma^2}\\ &= \frac{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}{\sigma^2}.\end{aligned}$$

□

1.6 Les modes de convergence

On se donne une suite de variables aléatoires $X_1, X_2, \ldots$. Les modes de convergence suivant sont rangés du plus faible au plus fort.

On dit que la suite converge en loi vers la v.a. X si :

$$\forall \phi : \mathbb{R} \to \mathbb{R} \text{ continue et bornée}, \quad \mathbb{E}[\phi(X_n)] \xrightarrow[n\to\infty]{} \mathbb{E}[\phi(X)],$$

ce qui est équivalent par le théorème porte-manteau à :

$$\forall t \in \mathbb{R} \text{ tel que } F_X \text{ est continue en } t, \quad F_{X_n}(t) \xrightarrow[n\to\infty]{} F_X(t).$$

On dit que la suite converge en probabilité vers la v.a. X si :

$$\forall \varepsilon > 0, \quad \mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0.$$

On dit que la suite converge L^1 vers la v.a. X si :

$$\mathbb{E}[|X_n - X|] \xrightarrow[n\to\infty]{} 0.$$

On dit que la suite converge L^2 vers la v.a. X si :

$$\mathbb{E}\left[|X_n - X|^2\right] \xrightarrow[n\to\infty]{} 0.$$

On dit que la suite converge vers la v.a. X presque sûrement si :

$$\mathbb{P}\left(\left\{\omega : X_n(\omega) \xrightarrow[n\to\infty]{} X(\omega)\right\}\right) = 1.$$

Toutes ces définitions s'étendent au cas de vecteurs aléatoires.

Les relations entre les convergences

La convergence presque sûre implique la convergence en probabilité. La convergence L^2 implique la convergence L^1. La convergence L^1 implique la convergence en probabilité. La convergence en probabilité implique la convergence en loi.

Les théorèmes *limite*

Lemme de l'application continue Soient $X_1, X_2, \dots$ une suite de v.a. qui converge presque sûrement (resp. en probabilité, en loi) vers X, et ϕ une application $\mathbb{R} \to \mathbb{R}$ continue. Alors $\phi(X_n)$ converge presque sûrement (resp. en probabilité, en loi) vers $\phi(X)$.

Loi forte des grands nombres. Soient $X_1, X_2, \dots$ une suite de v.a. i.i.d. qui sont L^1 (c'est-à-dire que $\mathbb{E}[|X_1|] < \infty$) dont on note $m = \mathbb{E}[X_1]$. Alors :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{p.s.} m.$$

Loi faible des grands nombres. Soient $X_1, X_2, \dots$ une suite de v.a. i.i.d. qui sont L^1 (c'est-à-dire que $\mathbb{E}[|X_1|] < \infty$) dont on note $m = \mathbb{E}[X_1]$. Alors :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\mathbb{P}} m.$$

Théorème de la limite centrale. Soient $X_1, X_2, \dots$ une suite de v.a. i.i.d. qui sont L^2 (c'est-à-dire que $\mathbb{E}\left[|X_1|^2\right] < \infty$) dont on note $m = \mathbb{E}[X_1]$ et $\sigma^2 = \mathbb{V}(X_1)$. Alors :

$$\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1) \Leftrightarrow \sqrt{n}\left(\overline{X}_n - m\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\sigma \sim \mathcal{N}\left(0, \sigma^2\right)$$

(l'équivalence s'obtient avec le LAC : poser $g : x \mapsto \sigma x$ pour passer de gauche à droite par exemple). Cela peut se réécrire avec les fonctions de répartition (Φ représente la cdf d'une loi normale centrée réduite) grâce au théorème portemanteau :

$$\forall t \in \mathbb{R},\ \mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - m}{\sigma} \leq t\right) \xrightarrow[n\to\infty]{} \Phi(t).$$

Théorème de la limite centrale multidimensionnel. Soient $X_1, X_2, \dots$ une suite de v.a. i.i.d. qui sont L^2 (c'est-à-dire que $\mathbb{E}\left[\|X_1\|^2\right] < \infty$) dont on note $m = \mathbb{E}[X_1]$ et Σ la matrice de variance-covariance de X_1. Alors :

$$\sqrt{n}\left(\overline{X}_n - m\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Lemme sans nom Soient $X_1, X_2, \dots$ une suite de v.a. qui converge en loi vers une constante c. Alors la convergence a également lieu en probabilité.

Autre lemme sans nom Soient $X_1, X_2, \dots$ une suite de v.a. qui converge presque sûrement (resp. en probabilité, L^1, L^2) vers X et $Y_1, Y_2, \dots$ une suite de v.a. qui converge presque sûrement (resp. en probabilité, L^1, L^2) vers Y. Alors $(X_1, Y_1), (X_2, Y_2), \dots$ converge presque sûrement (resp. en probabilité, L^1, L^2) vers (X, Y). La réciproque est vraie. Le résultat s'étend à un nombre quelconque fini de suites. **Attention** ce résultat n'est pas vrai pour la convergence en loi ! Voir le résultat suivant.

Lemme de Slutsky Soient $X_1, X_2, \dots$ une suite de v.a. qui converge en loi vers X et $Y_1, Y_2, \dots$ une suite de v.a. qui converge en loi vers une constante c. Alors $(X_1, Y_1), (X_2, Y_2), \dots$ converge en loi vers (X, c).

Corollaire du lemme de Slutsky Soient $X_1, X_2, \ldots$ une suite de v.a. qui converge en loi vers X et $Y_1, Y_2, \ldots$ une suite de v.a. qui converge en loi vers une constante c. Alors $X_n + Y_n$, $X_n - Y_n$, $X_n Y_n$ convergent en loi vers $X + c$, $X - c$, cX et, si $Y_n \neq 0$ p.s. et $c \neq 0$, X_n / Y_n converge en loi vers X/c. *Preuve :* combiner le lemme de Slutsky et le lemme de l'application continue.

Chapitre 2

Modèle statistique, estimateurs

2.1 Modèle statistique

2.1.1 Modèle et statistique sur un modèle

Un modèle statistique est utilisé pour modéliser des observations venant d'un phénomène aléatoire. Par rapport à un espace probabilisé, on ne connaît pas la loi ayant servi à produire les observations mais un ensemble de lois comprenant cette loi d'origine.

Définition 6. *Un modèle statistique* $(\mathbb{X}, \mathfrak{F}, \mathfrak{P})$ *est la donnée de :*

- *L'espace* $\mathbb{X}$ *des observations.*
- *Une tribu* $\mathfrak{F}$ *sur* $\mathbb{X}$ *représentant l'ensemble des évènements sur* $\mathbb{X}$*.*
- *Une famille* $\mathcal{P}$ *de mesures de probabilités sur* $\mathfrak{F}$*.*

Un modèle statistique est une famille d'expériences aléatoires ayant même espace d'observations et même tribu d'évènements. Une observation $x \in \mathbb{X}$ supposée être une réalisation d'une variable aléatoire $X : (\Omega, \mathcal{A}) \to (\mathbb{X}, \mathfrak{F})$ qui suit l'une des lois de probabilité (inconnue) de la famille $\mathfrak{P}$. C'est-à-dire que formellement on suppose qu'il existe $\omega \in \Omega$ tel que $x = X(\omega)$. L'objectif de l'inférence statistique est de déterminer cette loi de probabilité ou des caractéristiques de cette loi à partir de l'observation x.

On rappelle qu'on note $X \sim P$ pour indiquer que X suit la loi $P \in \mathfrak{P}$.

Exemple : épreuve de Bernoulli.

Soit X une va de loi de Bernoulli $\mathcal{B}(\theta)$, où θ est inconnu. Soit $x \in \{0, 1\}$ le résultat d'un lancer à pile ou face ; x est une réalisation de X. Un choix raisonnable de la famille $\mathcal{P}$ est l'ensemble des lois de Bernoulli de paramètre θ. L'espace des observations est $\mathbb{X} = \{0, 1\}$ et $\mathfrak{F}$ est, par exemple, l'ensemble des parties de $\mathbb{X}$. Au final le modèle s'écrit donc

$$(\{0, 1\}, \mathcal{P}(\{0, 1\}), \{\mathcal{B}(\theta), \theta \in]0, 1[\}) .$$

Exemple : loi uniforme.

Soit X une va de loi uniforme sur $[0, \theta]$, où θ est inconnu. Soit x une réalisation de X ; si $x = 0.1$, on peut en déduire que $\theta \geq 0.1$. Si $x = 2$, on peut en déduire que $\theta \geq 2$. Le but est bien sûr d'aller plus loin. Un choix raisonnable de la famille $\mathfrak{P}$ est l'ensemble des lois uniforme sur $[0, \theta]$. L'espace des observations est $\mathbb{X} = \mathbb{R}_+$ ou $\mathbb{X} = \mathbb{R}$ et $\mathfrak{F}$ est, par exemple, l'ensemble des boréliens de $\mathbb{X}$. Le modèle s'écrit donc par exemple

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{U}([0, \theta]), \theta > 0\}) .$$

En général on va observer non pas une mais n réalisations $x_1, ..., x_n$ d'une v.a. X, de manière indépendante. Soient $X_1, ..., X_n$ les v.a.i.i.d. sous-jacentes à ces observations. Les X_i sont à valeurs dans $(\mathbb{X}, \mathfrak{F})$. Soit P la loi commune des X_i. La loi du n-échantillon $(X_1, ..., X_n)$ est $P^{\otimes n}$ par indépendance et identique distribution des X_i.

Définition 7. *À tout modèle statistique* $(\mathbb{X}, \mathfrak{F}, \mathfrak{P})$ *et tout* $n \in \mathbb{N}^*$ *on associe le modèle d'échantillonnage associé au* n*-échantillon, le triplet :*

$$\left(\mathbb{X}^n, \mathfrak{F}^{\otimes n}, \left\{P^{\otimes n}, P \in \mathfrak{P}\right\}\right). \tag{2.1}$$

En particulier un modèle d'échantillonnage est lui-même un modèle statistique. Dans la pratique on a presque tout le temps plusieurs mesures indépendantes d'une même quantité donc on travaille presque toujours avec un modèle d'échantillonnage.

Ce modèle traduit simplement le fait que dans une expérience répétée de façon indépendante, la loi de probabilité du n-échantillon est la loi produit et l'espace sous-jacent est le produit cartésien de l'espace associé à l'un des éléments de l'échantillon.

Exemple : répétition de n épreuves de Bernoulli indépendantes.

Le modèle d'échantillonnage associé est

$$\left(\{0,1\}^n, \mathfrak{F}^{\otimes n}, \left\{\mathcal{B}(\theta)^{\otimes n}, \theta \in [0,1]\right\}\right),$$

avec $\mathfrak{F} = \{\varnothing, \{0\}, \{1\}, \{0,1\}\}$.

Exemple : répétition de n épreuves d'une loi uniforme sur $[0, \theta]$.

Le modèle d'échantillonnage associé est

$$\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \left\{\mathcal{U}([0,\theta])^{\otimes n}, \theta > 0\right\}\right).$$

Définition 8. *Une statistique* T *sur un modèle* $(\mathbb{X}, \mathfrak{F}, \mathfrak{P})$ *et à valeurs dans* $(\mathbb{Y}, \mathfrak{T})$ *est une application mesurable*

$$T : (\mathbb{X}, \mathfrak{F}) \longrightarrow (\mathbb{Y}, \mathfrak{T}) \tag{2.2}$$

$$x \mapsto T(x) \tag{2.3}$$

Pour une réalisation x de X, $T(x)$ est la réalisation de la v.a. $T(X)$.

Une statistique est donc simplement une fonction (mesurable) des *seules* observations de l'expérience (et pas du paramètre à estimer). On la verra souvent comme une variable aléatoire en considérant $T(X)$. On rappelle qu'en général on travaillera sur des modèles d'échantillonnage et donc on considérera des statistiques définies sur des modèles d'échantillonnage, cf les exemples suivants.

Toute statistique étant une fonction mesurable des données, elle permet de définir une variable aléatoire à partir des données initiales. La loi image P_T de P par T permet alors de définir un modèle image à partir du modèle statistique initial, le modèle $(\mathbb{Y}, \mathfrak{T}, \{P_T : P \in \mathfrak{P}\})$.

Considérer $T = T(X)$ plutôt que X pour mener l'inférence consiste à travailler directement dans le modèle image par T :

$$\mathbb{P}(T \in B) = \mathbb{P}(T(X) \in B) = P_T(B) = \mathbb{P}(X \in T^{-1}(B)) = P\left(T^{-1}(B)\right), \ \forall P \in \mathfrak{P}, \forall X \sim P, \forall B \in \mathfrak{T} \tag{2.4}$$

L'information concernant la loi inconnue de X au travers de la statistique $T(X)$ est contenue dans la tribu $\sigma(T) \subset \sigma(X)$, et l'on a égalité si, et seulement si, T est bijective. Habituellement, $\sigma(T)$ est plus petite que $\sigma(X)$ car on attend d'une statistique qu'elle réduise la tribu et condense l'information. En effet :

$$\begin{aligned}\sigma(T) &= \left\{(T \circ X)^{-1}(B), B \in \mathfrak{T}\right\} \\ &= \left\{X^{-1}\left(T^{-1}(B)\right), B \in \mathfrak{T}\right\} \\ &\subseteq \left\{X^{-1}(F), F \in \mathfrak{F}\right\} = \sigma(X).\end{aligned}$$

Exemple : répétition de n épreuves de Bernoulli indépendantes.

$$S(x_1, x_2, ..., x_n) = \sum_{i=1}^{n} x_i \tag{2.5}$$

Définit une statistique qui est la somme de toutes les observations du n-échantillon. Dans l'exemple précédent du modèle de Bernoulli, on a donc $S(X) = \sum_{i=1}^{n} X_i$ et la loi image est la loi binomiale $P_S = \mathcal{B}(n, \theta)$. Le modèle image est donné par

$$(\llbracket 0, n \rrbracket, \mathcal{P}(\llbracket 0, n \rrbracket), \{\mathcal{B}(n, \theta), \theta \in]0, 1[\}) .$$

Exemple : statistique définie par le maximum d'une loi uniforme sur $[0, \theta]$.

$$T(X_1, X_2, ..., X_n) = \max_{1 \leq i \leq n} X_i \tag{2.6}$$

2.1.2 Quelques types de modèles statistiques

Rappel : mesure dominée et théorème de Radon-Nikodym

Définition 9. *Une mesure μ sur $(\mathbb{X}, \mathfrak{F})$ est σ-finie si il existe une suite $E_1, E_2, \cdots \in \mathfrak{F}$ tels que $\mu(E_i) < \infty$ et $\mathbb{X} = \bigcup_{i=1}^{\infty} E_i$.*

Toute mesure finie est σ-finie, en particulier les mesures de probabilité. La mesure de Lebesgue aussi est σ-finie. La mesure de comptage d'un ensemble dénombrable aussi.

Étant donné deux mesures σ-finies μ et ν sur l'espace $(\mathbb{X}, \mathfrak{F})$, on dit que ν est absolument continue par rapport μ et l'on note $\nu \ll \mu$, si

$$\forall B \in \mathfrak{F}, \mu(B) = 0 \Rightarrow \nu(B) = 0. \tag{2.7}$$

On rappelle qu'en ce cas, le théorème de Radon-Nikodym assure que ν possède une densité f par rapport à μ définie par

$$\nu(B) = \int_B f(x) \mathrm{d}\mu(x) \tag{2.8}$$

On note $f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$. f est positive, mesurable et s'appelle densité ou dérivée au sens de Radon-Nikodyn de ν par rapport à μ. Elle est unique μ-p.p.

Modèle dominé

Définition 10. *Un modèle statistique $(\mathbb{X}, \mathfrak{F}, \mathfrak{P})$ est dominé si, et seulement si, tout $P \in \mathfrak{P}$ est dominé par une même mesure μ σ-finie. μ est la mesure dominante du modèle.*

Si un modèle est dominé par une mesure σ-finie μ, alors le modèle image est dominé par la mesure μ_S définie par $\mu_S(B) = \mu\left(S^{-1}(B)\right)$. En effet, pour tout ensemble mesurable B tel que $\mu_S(B) = 0$, $\mu\left(S^{-1}(B)\right) = 0$ donc $P(S^{-1}(B)) = 0$ par domination donc $P_S(B) = 0$.

Modèle paramétrique

Définition 11. *Un modèle statistique $(\mathbb{X}, \mathfrak{F}, \mathfrak{P})$ est paramétré si les éléments de $\mathfrak{P}$ peuvent être décrit par un paramètre. Cela veut dire qu'on peut écrire $\mathfrak{P} = \{P_\theta, \theta \in \Theta\}$.*

Si $\theta \mapsto P_\theta$ est injective (donc bijective) on dit que le modèle (muni de cette paramétrisation) est identifiable (cela veut dire que $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$).

Le modèle est paramétrique si $\Theta \subset \mathbb{R}^p$ pour un entier p donné. Sinon, il est non-paramétrique.

À noter que tout modèle peut être paramétré par $\mathfrak{P} = \{P, P \in \mathfrak{P}\}$ et cette paramétrisation est identifiable (mais pas très utile).

Exemple : épreuve de Bernoulli.

X suit de loi de Bernoulli $\mathcal{B}(\theta)$ paramétrée par $\theta \in [0,1]$. Le modèle est donc paramétrique, identifiable car à chaque valeur de θ correspond une unique loi de Bernoulli. Ce modèle est également dominé par la mesure de comptage discrète sur $\mathbb{N}$.

Si, au lieu de considérer la paramétrisation du modèle $\{\mathcal{B}(\theta), \theta \in [0,1]\}$, on considère $\{\mathcal{B}(\cos^2\theta), \theta \in \mathbb{R}\}$, le modèle n'est plus identifiable car $P_\theta = P_{\theta+\pi}$.

Exemple : modèle d'échantillonnage gaussien.

On considère un n-échantillon de loi gaussienne de moyenne m et écart type σ. Le modèle statistique correspondant

$$\left(\mathbb{R}^n, B(\mathbb{R})^{\otimes n}, \left\{\mathcal{N}(m,\sigma^2)^{\otimes n} : m \in \mathbb{R}, \sigma^2 > 0\right\}\right) \tag{2.9}$$

est paramétrique, dominé et identifiable. Il est paramétré par le couple (m,σ^2), dominé par la mesure de Lebesgue sur $\mathbb{R}^n$ et identifiable car toute loi gaussienne est caractérisée par sa moyenne et sa variance. Si on considère la paramétrisation $(m,\sigma) \in \mathbb{R}\times\mathbb{R}^*$ au lieu de $(m,\sigma^2) \in \mathbb{R}\times\mathbb{R}_+^*$, le modèle n'est plus identifiable.

Le modèle

$$\left(\mathbb{R}^n, B(\mathbb{R})^{\otimes n}, \left\{\mathcal{N}(m,\sigma^2)^{\otimes n} : m \in \mathbb{R}, \sigma^2 \geq 0\right\}\right) \tag{2.10}$$

est paramétrique et identifiable mais pas dominé par la mesure de Lebesgue car il contient toutes les lois $\mathcal{N}(m,0) = \delta_m$.

Exemple : modèles non paramétriques.

L'ensemble des lois de probabilités sur $\mathbb{R}$ n'est pas paramétrique. Il en est de même de l'ensemble des lois qui ont une densité par rapport à la mesure de Lebesgue ou bien l'ensemble des lois réelles continues dont la densité est symétrique. De façon générale, lorsque le paramètre n'appartient pas à un espace vectoriel de dimension finie, on parle de modèle non paramétrique.

Exemple : modèles semi-paramétriques.

On peut s'intéresser à des problèmes dépendant d'un paramètre $\theta \in \mathbb{R}^p$, mais également d'un autre paramètre appartenant à un espace de dimension infinie. Ce second paramètre représente souvent un terme de nuisance. On parle en ce cas de problème semi-paramétrique.

Vraisemblance

Définition 12. *On considère un modèle statistique $(\mathbb{X}, \mathfrak{F}, \{P_\theta; \theta \in \Theta\})$ dominé par une mesure μ σ-finie. Et un objet aléatoire X tiré selon une loi du modèle P_{θ^*}. La (une) vraisemblance de X est une fonction $\theta \mapsto L(\theta; X)$ définie par*

$$L(\theta; X) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(X) \tag{2.11}$$

avec $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$ une densité de P_θ par rapport à μ. Si le modèle est un modèle d'échantillonnage on parle de vraisemblance de l'échantillon X.

Remarque : $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$ n'est pas unique, mais elle est unique à un ensemble μ-négligeable près, donc, vu que $\mathcal{L}(X) = P_{\theta^*} \ll \mu$, $L(\theta; X)$ est unique presque sûrement.

On rappelle que si $\mathbb{X} = \mathbb{R}$ et $\mu = \lambda$ la mesure de Lebesgue, $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$ correspond à la notion classique de densité d'une variable aléatoire absolument continue. On rappelle aussi que si $\mathbb{X} = \mathbb{R}$ et $\mu =$ la mesure de comptage d'un ensemble dénombrable $\mathbb{F}$, $\frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$ correspond à la fonction de masse de probabilité d'une variable aléatoire discrète à support inclus dans $\mathbb{F}$.

Remarque : la vraisemblance est donc une fonction aléatoire de θ à X fixé. À ne pas confondre avec une densité qui est une fonction de $x \in \mathbb{X}$ à θ fixé !! $\theta \mapsto L(\theta; X)$ est une vraisemblance, $x \mapsto L(\theta; x)$ est une densité.

Si on se place dans un modèle d'échantillonnage (presque toujours le cas en pratique) $\left(\mathbb{X}^n, \mathfrak{F}^{\otimes n}, \left\{P_\theta^{\otimes n}; \theta \in \Theta\right\}\right)$ associé à $(\mathbb{X}, \mathfrak{F}, \{P_\theta; \theta \in \Theta\})$ dominé par μ, alors une densité de $P_\theta^{\otimes n}$

par rapport à $\mu^{\otimes n}$ est $(x_1, \dots x_n) \mapsto L_n(\theta; x_1, \dots x_n) = \prod_{i=1}^n \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(x_i)$ et donc la vraisemblance d'un échantillon $X = (X_1, \dots, X_n)$ est la fonction

$$\theta \mapsto L_n(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}(X_i) \tag{2.12}$$

En pratique on la notera souvent juste $L_n(\theta)$ en omettant la dépendance en $(X_1, \dots, X_n)$.

Exemple : échantillon uniforme sur $[0, \theta]$.

$$L_n(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \left(\frac{1}{\theta} \mathbb{1}_{[0,\theta]}(X_i) \right) = \frac{1}{\theta^n} \prod_{i=1}^n \mathbb{1}_{[0,\theta]}(X_i) = \frac{1}{\theta^n} \mathbb{1}_{[0,\theta]^n}(X) \tag{2.13}$$

Où $X = (X_1, \dots, X_n)$ et la dernière indicatrice vaut 1 si, et seulement si, toutes les coordonnées X_i du vecteur X appartiennent à $[0, \theta]$.

Exemple : échantillon de Bernoulli de paramètre θ.

$$L_n(\theta; X_1, \dots, X_n) = \prod_{i=1}^n \left(\theta^{X_i} (1-\theta)^{1-X_i} \mathbb{1}_{\{0,1\}}(x_i) \right) = \theta^S (1-\theta)^{n-S} \mathbb{1}_{\{0,1\}^n}(X) \tag{2.14}$$

où $X = (X_1, \dots, X_n)$, $S = \sum_{i=1}^n X_i$ et la dernière indicatrice vaut 1 si $X_i = 0$ ou 1 pour tout i, donc en fait elle vaut 1 p.s. car la loi de X est dans le modèle d'échantillonnage. Vu que de toute façon toutes les égalités qu'on écrit sont vraies seulement presque sûrement, on a finalement $L_n(\theta; X_1, \dots, X_n) = \theta^S (1-\theta)^{n-S}$.

Modèle homogène

Définition 13. *Un modèle paramétrique est homogène si le support de toutes lois P_θ est le même.*

Cela signifie également que le support ne dépend pas de θ, ou encore que toutes les lois de $\mathcal{P}$ sont équivalentes, c'est-à-dire :

$$\forall \theta, \theta', \ P_\theta \ll P_{\theta'} \tag{2.15}$$

Homogène implique dominé (par n'importe quelle loi du modèle) et dans un modèle homogène, les ensembles négligeables sont les mêmes pour toutes les mesures P_θ. Enfin, un modèle est homogène si, et seulement si,

$$\exists\ \nu\ \sigma\text{-finie} : \ \forall \theta, \ P_\theta \ll \nu \text{ et } \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu} > 0 \ \nu - \text{p.p.} \tag{2.16}$$

2.2 Estimateur, biais et erreur quadratique

On se place dans un modèle statistique paramétrique $(\mathbb{X}, \mathfrak{F}, P)$ avec $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$ et l'on considère une variable aléatoire $X \sim P_\theta$.

Définition 14. *Un estimateur T de θ est une statistique à valeurs dans un sur-ensemble de Θ. En général on identifie T et $T(X)$. Et dans le cadre d'un modèle d'échantillonnage de taille n, on note plutôt l'estimateur $\hat{\theta}_n$.*

Un estimateur doit être une fonction des seules données et ne pas dépendre du paramètre θ qu'il est censé estimer. On attend bien sûr d'un estimateur qu'il donne une valeur proche de la vraie valeur de θ.

Définition 15. *Le biais d'un estimateur T est la quantité*

$$\mathcal{B}(T, \theta) = \mathbb{E}_\theta[T(X)] - \theta. \tag{2.17}$$

Le risque quadratique moyen (ou erreur quadratique moyenne) de l'estimateur est

$$\mathcal{R}(T, \theta) = \mathbb{E}_\theta[(T(X) - \theta)^2] \tag{2.18}$$

$$= \mathbb{V}_\theta(T(X)) + \mathcal{B}(T, \theta)^2. \tag{2.19}$$

Dans toute la suite, les notations $\mathbb{E}_\theta, \mathbb{V}_\theta, \mathbb{P}_\theta$ sont de légers abus de notation pour signaler que $X \sim P_\theta$ (on voit le conflit de notation et donc le caractère abusif de la notation). L'égalité (2.19) *s'appelle la décomposition biais-variance, on la démontre ci-dessous.*

Preuve de (2.19)

$$\begin{aligned}\mathbb{E}_\theta[(T(X)-\theta)^2] &= \mathbb{E}_\theta[(T(X)-\mathbb{E}_\theta[T(X)]+\mathbb{E}_\theta[T(X)]-\theta)^2] \\ &= \mathbb{E}_\theta[(T(X)-\mathbb{E}_\theta[T(X)])^2 + 2(T(X)-\mathbb{E}_\theta[T(X)])(\mathbb{E}_\theta[T(X)]-\theta) + (\mathbb{E}_\theta[T(X)]-\theta)^2] \\ &= \mathbb{V}_\theta(T(X)) + 2(\mathbb{E}_\theta[T(X)]-\theta)\mathbb{E}_\theta[T(X)-\mathbb{E}_\theta[T(X)]] + \mathcal{B}(T,\theta)^2 \\ &= \mathbb{V}_\theta(T(X)) + \mathcal{B}(T,\theta)^2.\end{aligned}$$

Si $\mathcal{B}(T,\theta) \neq 0$, on dit que l'estimateur est biaisé, sinon il est dit non-biaisé.

La qualité d'un estimateur se juge à son risque quadratique : plus il est petit, meilleur est l'estimateur. Pour comparer deux estimateurs on compare donc leur risque.

Exemple :

Considérons un n-échantillon $X_1, ..., X_n$ d'une v.a. X d'espérance m et de variance σ^2, sa moyenne empirique $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et sa variance empirique :

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2. \tag{2.20}$$

$\overline{X}_n$ et $\widehat{\sigma}_n^2$ sont des estimateurs respectifs de la moyenne m et de la variance σ^2 de X.

On voit facilement que $\mathbb{E}\left[\overline{X}_n\right] = m$ et un calcul rapide (fait en cours, utilisant d'abord que $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\overline{X}_n\right)^2$ puis la linéarité de l'espérance) donne $\mathbb{E}\left[\widehat{\sigma}_n^2\right] = \frac{n-1}{n}\sigma^2$.

$\overline{X}_n$ est donc un estimateur sans biais de m tandis que $\widehat{\sigma}_n^2$ est un estimateur biaisé de la variance. On peut cependant le débiaiser en posant $S_n^2 = \frac{n}{n-1}\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. (Et surtout pas en posant $\widehat{\sigma}_n^2 + \frac{\sigma^2}{n}$ qui dépend de σ^2 inconnu et qui est donc incalculable !)

Estimateur par substitution

Définition 16. *Soit $T : \mathbb{X} \to \Theta$ un estimateur de θ et $\psi : \Theta \longrightarrow \mathbb{R}^d$ une fonction mesurable. Un estimateur par substitution de $\psi(\theta)$ est un estimateur de la forme $\psi \circ T$.*

Estimateur par la méthode des moments

Étant donné une variabe aléatoire X tirée selon une loi du modèle P_θ, on note $m^{(k)}(\theta) = \mathbb{E}\left[X^k\right]$ son moment d'ordre k s'il existe (noter qu'on écrit la dépendance en θ du moment) et

$$\widehat{m}_n^{(k)} = \frac{1}{n}\sum_{i=1}^n X_i^k \tag{2.21}$$

le moment empirique d'ordre k d'un n-échantillon $(X_1, \ldots, X_n)$.

Définition 17. *Un estimateur par la méthode des moments de θ est obtenu en inversant la relation entre θ et $m^{(1)}(\theta), \ldots, m^{(k)}(\theta)$ pour un certain k, puis en remplaçant $m^{(1)}(\theta), \ldots, m^{(k)}(\theta)$ par $\widehat{m}_n^{(1)}, \ldots, \widehat{m}_n^{(k)}$. C'est-à-dire, que cela consiste en trouver $k \geq 1$ et une fonction f tels que $\theta = f\left(m^{(1)}(\theta), \ldots, m^{(k)}(\theta)\right)$, puis à poser $\hat{\theta}_n = f\left(\widehat{m}_n^{(1)}, \ldots, \widehat{m}_n^{(k)}\right)$. C'est donc un certain type d'estimateur par substitution.*

Remarque : il n'y a pas unicité de l'entier k ni de l'estimateur, on parle donc d'"un" estimateur.

Remarque : le principe s'étend pour l'estimation d'un $\psi(\theta)$, et revient à estimer d'abord θ par $\hat{\theta}_n$ par la méthode des moments puis effectuer la substitution et estimer donc enfin $\psi(\theta)$ par $\psi(\hat{\theta}_n)$.

Remarque : parfois il est plus pratique d'utiliser les moments centrés plutôt que les moments. Dans ce cas-là on utilisera donc les moments centrés empiriques d'ordre k définis par $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^k$.

Exemple :

L'estimateur des moments d'un n-échantillon de loi L^2 pour le paramètre (m, σ^2) est donc donné par le couple $(\overline{X}_n, \hat{\sigma}_n^2)$. On retrouve les estimateurs proposés précédemment.

Exemple :

Considérons un n-échantillon $(X_1, ..., X_n)$ de loi commune une loi de Poisson de paramètre θ inconnu. On sait que $\mathbb{E}_{X\sim\mathcal{P}(\theta)}[X] = \theta$. On va donc poser comme estimateur des moments de θ la statistique $\overline{X}_n = \widehat{m}_n^{(1)}$.

Exemple :

Considérons un n-échantillon $(X_1, ..., X_n)$ de loi commune une loi exponentielle de paramètre θ inconnu (la densité de cette loi est $x \mapsto \theta\exp(-\theta x)\,\mathbb{1}_{x>0}$). On sait que $\mathbb{E}_{\mathcal{E}(\theta)}[X] = 1/\theta$. On va donc poser comme estimateur des moments de θ la statistique T_n telle que :

$$\overline{X}_n = \widehat{m}_n^{(1)} = \frac{1}{T_n} \Rightarrow T_n = \frac{1}{\overline{X}_n}. \tag{2.22}$$

On sait aussi que $\mathbb{V}_{X\sim\mathcal{E}(\theta)}(X) = 1/\theta^2$. On peut donc également estimer θ par :

$$\frac{1}{\sqrt{\hat{\sigma}_n^2}} = \frac{1}{\sqrt{\widehat{m}_n^{(2)} - \left(\widehat{m}_n^{(1)}\right)^2}}. \tag{2.23}$$

On voit bien la non-unicité de la méthode.

Exemple :

L'estimateur des moments d'un n-échantillon de loi uniforme sur $[0, \theta]$ où θ est le paramètre inconnu est donné par la statistique $T_n = 2\overline{X}_n$ (à vérifier).

Estimateur du maximum de vraisemblance

Définition, propriétés et exemples

Définition 18. *Un estimateur du maximum de vraisemblance de θ (on notera e.m.v.) est, sous réserve d'existence, un élément $\widehat{\theta}$ tel que :*

$$\widehat{\theta} \in \arg\max_{\theta\in\Theta} L(\theta; X) \tag{2.24}$$

Ce maximum peut ne pas exister, n'être pas unique, ou être difficile à calculer.

Théorème 2 (Une condition suffisante d'existence de l'e.m.v.)**.** *Si Θ est un espace compact et si L est continue sur Θ, alors un estimateur du maximum de vraisemblance existe.*

Déterminer le maximum de vraisemblance est équivalent à déterminer le maximum de la log-vraisemblance (on notera cette log-vraisemblance $\ln L(\theta; X) = \ell(\theta; X)$). Si L est différentiable, l'estimateur du maximum de vraisemblance est alors nécessairement solution de l'équation

$$\frac{\partial}{\partial\theta}\log L(\theta; X) = \frac{\partial}{\partial\theta}\ell(\theta; X) = \nabla_\theta\ell(\theta; X) = 0. \tag{2.25}$$

Un θ_c qui vérifie (2.25) est appelé un point critique.

La dérivée $\nabla_\theta\ell$ de la log-vraisemblance (ou son gradient dans le cas multidimensionnel) s'appelle le score.

L'équation précédente donne une condition nécessaire, mais non suffisante, pour être un maximum local. On rappelle qu'un point $\hat{\theta}$ est un maximum est local s'il existe un voisinage de $\widehat{\theta}$ dans Θ tel que

$$L(\theta; X) \leq L(\widehat{\theta}; X) \tag{2.26}$$

pour tout θ dans ce voisinage. Un maximum est global si l'inégalité précédente est vérifiée pour tout $\theta \in \Theta$.

Dans le cas où Θ est un intervalle de $\mathbb{R}$ (le paramètre à estimer est scalaire), une condition suffisante d'existence et d'unicité locale est que $\theta \mapsto L(\theta; X)$ soit une fonction de classe C^2 sur Θ (on rappelle que X est fixée) et que les deux conditions suivantes soient réalisées :

$$\begin{cases} \frac{\mathrm{d}\ell}{\mathrm{d}\theta}(\theta; X) = 0 \\ \\ \frac{\mathrm{d}^2\ell}{\mathrm{d}\theta^2}(\theta; X) < 0. \end{cases} \tag{2.27}$$

La solution doit par ailleurs appartenir à l'intérieur $\mathring{\Theta}$ de Θ. Les conditions assurent l'existence et l'unicité d'un maximum local uniquement et la valeur θ_c qui réalise ce maximum n'est donc pas forcément l'estimateur du maximum de vraisemblance (car il peut ne pas être un maximum global). Par contre, si de plus θ_c est l'unique point critique, alors il réalise bien un maximum global. Dans la pratique, c'est comme cela que l'on démontre qu'un estimateur est un e.m.v. quand ℓ est suffisamment régulière.

Remarque : on a existence et unicité d'un maximum global si la fonction ℓ est strictement concave sur Θ, mais c'est rare dans la pratique. On préférera toujours la méthode décrite plus haut.

Le cas multidimensionnel généralise ce qui précède mais évidemment avec le gradient qui remplace la dérivée et la hessienne qui remplace la dérivée seconde. Il se résume dans le théorème suivant.

Théorème 3. *Sous les trois hypothèses de régularité suivantes :*

- *Le modèle est identifiable et homogène.*
- Θ *est un ouvert ou un compact de* $\mathbb{R}^p$.
- *La fonction* $\theta \mapsto L(\theta; X)$ *est de classe* C^2 *sur* Θ.

Alors un point θ_c *qui vérifie*

$$\begin{cases} \nabla_\theta \ell(\theta_c; X) = 0 \\ H_\ell(\theta_c; X) \prec 0. \end{cases} \tag{2.28}$$

réalise un maximum local de L *(et il est unique localement). Ici* $\ell(\theta; X) = \ln L(\theta; X)$ *est la log-vraisemblance,* $\nabla_\theta \ell$ *le gradient de* ℓ *en* θ *et* H_ℓ *est la matrice hessienne de* ℓ *en* θ.

Rappel : le gradient est le vecteur (identifié à la matrice colonne) des dérivées partielles d'ordre 1 et la matrice hessienne est la matrice des dérivées partielles secondes. Elle est symétrique.

$$\nabla_\theta \ell(\theta; X) = \left(\frac{\partial}{\partial \theta_1} \ell(\theta; X), \ldots, \frac{\partial}{\partial \theta_p} \ell(\theta; X) \right) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ell(\theta; X) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ell(\theta; X) \end{pmatrix}$$

et

$$H_\ell(\theta; X) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell(\theta; X) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ell(\theta; X) & \ldots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \ell(\theta; X) \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ell(\theta; X) & \frac{\partial^2}{\partial \theta_2^2} \ell(\theta; X) & \ldots & \frac{\partial^2}{\partial \theta_2 \partial \theta_p} \ell(\theta; X) \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \ell(\theta; X) & \ldots & & \frac{\partial^2}{\partial \theta_p^2} \ell(\theta; X) \end{pmatrix}.$$

Remarque : La condition $H_\ell(\theta_c; X) \prec 0$ signifie que la matrice hessienne est définie négative en θ_c. C'est l'analogue multi-dimmensionnel de la condition $\frac{\mathrm{d}^2\ell}{\mathrm{d}\theta^2}(\theta_c; X) < 0$. On montre qu'une matrice symétrique est définie négative en montrant que ses valeurs propres sont toutes < 0.

Remarque importante : Le théorème n'assure pas que le point critique θ_c est un e.m.v. Comme dans le cas uni-dimmensionnel, par contre si c'est l'unique point critique alors c'est bien un e.m.v.

Remarque : également comme dans le cas uni-dimmensionnel, on a existence et unicité d'un maximum global si la fonction ℓ est strictement concave sur Θ, mais c'est rare dans la pratique. On préférera toujours la méthode décrite plus haut.

Exemple :

Soit $X = (X_1, ..., X_n)$ un échantillon gaussien $\mathcal{N}(m, \sigma^2)$ et $\theta = (m, \sigma^2)$, $m \in \mathbb{R}$, $\sigma^2 > 0$.

Le modèle est paramétrique, dominé par la mesure de Lebesgue et la vraisemblance associée est

$$L(\theta; X) = \prod_{i=1}^{n} f_\theta(X_i) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - m)^2\right) \tag{2.29}$$

Le support du modèle est $\mathbb{R}^n$; celui-ci est donc homogène et la vraisemblance est clairement de classe C^2 en θ. Par conséquent, l'e.m.v. est solution de l'équation de vraisemblance

$$\nabla_\theta \ell(\theta; X) = 0. \tag{2.30}$$

La solution θ_c réalise un maximum local si la matrice hessienne $H_\ell(\theta_c; X)$ est définie négative.

$$\ell(\theta; X) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 \tag{2.31}$$

$$\nabla_\theta \ell(\theta; X) = \left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)\ ;\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - m)^2\right) \tag{2.32}$$

Ce gradient s'annule si, et seulement si

$$\sum_{i=1}^{n}(X_i - m) = 0 \text{ et } \sum_{i=1}^{n}(X_i - m)^2 = n\sigma^2 \tag{2.33}$$

$$\iff m = \overline{X}_n \text{ et } \sigma^2 = \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \tag{2.34}$$

Un point critique de ℓ est donc $\widehat{\theta}_n = \left(\overline{X}_n, \hat{\sigma}_n^2\right)$ et il est unique. La matrice hessienne de ℓ en ce point est (à vérifier à titre d'exercice)

$$H_\ell(\widehat{\theta}_n) = \begin{pmatrix} -n/\hat{\sigma}_n^2 & 0 \\ 0 & -n/(2\hat{\sigma}_n^4) \end{pmatrix} \tag{2.35}$$

qui est clairement définie négative. Par unicité du point critique, $\widehat{\theta}_n$ est l'e.m.v. Pour vérifier les calculs précédents, on aura intérêt à poser $v = \sigma^2$ pour éviter les erreurs au moment de la dérivation en σ^2.

Exemple :

Soit $X = (X_1, ..., X_n)$ un n-échantillon de loi uniforme sur $[0, \theta]$ où $\theta > 0$.

Le modèle est paramétrique, dominé par la mesure de Lebesgue sur $\mathbb{R}^n$. La vraisemblance de l'échantillon est

$$L(\theta; X) = \prod_{i=1}^{n}\left(\frac{1}{\theta}\mathbb{1}_{[0,\theta]}(X_i)\right) = \frac{1}{\theta^n}\mathbb{1}_{[0,\theta]^n}(X) = \frac{1}{\theta^n}\mathbb{1}_{0 \le X_{(1)} \le X_{(n)} \le \theta}(X) = \frac{1}{\theta^n}\mathbb{1}_{X_{(n)} \le \theta}(X) \tag{2.36}$$

où $X_{(1)}$ est le min des X_i et $X_{(n)}$ le max. Le support dépend de θ et le modèle n'est donc pas homogène. Nous ne pouvons pas déterminer le maximum de vraisemblance par la méthode précédente. On remarque par contre que, à X fixé, la fonction de vraisemblance est positive et décroissante en θ à partir de $X_{(n)}$ et, avant, elle est nulle. Le maximum est donc atteint en $\widehat{\theta} = X_{(n)}$. L'e.m.v. existe et est unique :

$$\widehat{\theta}_n = X_{(n)} = \max_{1 \le i \le n} X_i \tag{2.37}$$

On montre facilement que $X_{(n)}/\theta$ suit une loi beta de paramètres $(n,1)$, de densité $x \mapsto \frac{\Gamma(n+1)}{\Gamma(n)\Gamma(1)} x^{n-1}(1-x)^{1-1}\mathbb{1}_{[0,1]}(x) = nx^{n-1}\mathbb{1}_{[0,1]}(x)$ (cf l'annexe). La densité du maximum est donc

$$x \mapsto \frac{nx^{n-1}}{\theta^n}\mathbb{1}_{[0,\theta]}(x). \tag{2.38}$$

Ainsi, $\mathbb{E}[X_{(n)}] = \frac{n}{n+1}\theta$ et l'e.m.v. a donc un biais égal à $\mathcal{B}(\widehat{\theta}_n, \theta) = -\frac{\theta}{n+1}$.

Exemple :

Si $X_1, ..., X_n$ est un n-échantillon de loi uniforme sur $[\theta, \theta+1]$, on peut montrer (en exercice) que tout estimateur de θ compris entre $X_{(n)} - 1$ et $X_{(1)}$ est un estimateur du maximum de vraisemblance. Il n'y a alors pas unicité de cet e.m.v.

Reparamétrisation

Trouver un estimateur par substitution de l'estimateur du maximum de vraisemblance s'appelle une reparamétrisation. On montre que quelque soit l'application ψ, l'e.m.v. est invariant par reparamétrisation.

Théorème 4 (de Zehna)**.** *Soit $\widehat{\theta}$ un e.m.v. de θ et $\psi : \Theta \to \Xi$ une fonction quelconque. Alors un e.m.v. de $\psi(\theta)$ est $\psi(\widehat{\theta})$*

Démonstration :

Soit $\eta = \psi(\theta) \in \Xi$. Il faut déjà définir la vraisemblance $\widetilde{L}(\eta; X)$ de η. Si ψ est bijective, on pose très intuitivement

$$\widetilde{L}(\eta; X) = L(\psi^{-1}(\eta); X). \tag{2.39}$$

Sinon, on pose

$$\widetilde{L}(\eta; X) = \sup_{\tilde{\theta}:\psi(\tilde{\theta})=\eta} L(\tilde{\theta}; X). \tag{2.40}$$

Soit $\widehat{\eta}$ l'e.m.v. de η. Alors par définition, $\widehat{\eta}$ maximise la vraisemblance :

$$\widetilde{L}(\widehat{\eta}; X) = \sup_{\eta} \widetilde{L}(\eta; X) = \sup_{\eta} \sup_{\tilde{\theta}:\psi(\tilde{\theta})=\eta} L(\tilde{\theta}; X) = L(\widehat{\theta}; X) = \sup_{\tilde{\theta}:\psi(\tilde{\theta})=\psi(\hat{\theta})} L(\tilde{\theta}; X) = \widetilde{L}(\psi(\widehat{\theta}); X) \tag{2.41}$$

Donc $\psi(\widehat{\theta})$ maximise bien la vraisemblance d'η.

□

Chapitre 3

Propriétés asymptotiques

3.1 Propriétés asymptotiques et théorèmes limites

Considérons un estimateur T_n de θ dans un modèle d'échantillonnage. On dispose alors d'un échantillon $X_1, ..., X_n$ de v.a.i.i.d. de loi commune $\mathbb{P}_\theta$ et T_n est une fonction $T_n(X_1, ..., X_n)$ de $(X_1, ..., X_n)$, de sorte que T_n dépend de n. On dit que :

T_n est asymptotiquement sans biais si

$$\forall \theta \in \Theta, \ \lim_{n \to +\infty} \mathbb{E}_\theta[T_n] = \theta \tag{3.1}$$

T_n est fortement consistant pour θ si

$$\forall \theta \in \Theta, \ T_n \xrightarrow{\text{p.s.}} \theta \Longleftrightarrow \mathbb{P}_\theta \left[\lim_{n \to +\infty} T_n = \theta \right] = 1 \tag{3.2}$$

T_n converge en moyenne quadratique si

$$\forall \theta \in \Theta, \ T_n \xrightarrow{L^2} \theta \Longleftrightarrow \lim_{n \to +\infty} \mathbb{E}\left[|T_n - \theta|^2 \right] = \lim_{n \to +\infty} \mathcal{R}(\theta) = 0 \tag{3.3}$$

T_n est (faiblement) consistant pour θ si

$$\forall \theta \in \Theta, \ T_n \xrightarrow{\mathbb{P}} \theta \Longleftrightarrow \forall \epsilon > 0, \lim_{n \to +\infty} \mathbb{P}_\theta \left[|T_n - \theta| > \epsilon \right] = 0 \tag{3.4}$$

(Formellement il faut considérer une suite infinie $X_1, \ldots, X_n, \ldots$ définie pour Ω pour que tout ait un sens.)

La consistance forte implique la consistance. La convergence en moyenne quadratique est équivalente à la convergence du risque quadratique vers 0, et elle implique la consistance.

On dit que l'estimateur T_n est asymptotiquement normal si $\Theta \subseteq \mathbb{R}^p$ et si il existe une suite $(v_n)_n$ avec $v_n \geq 0$ et $v_n \underset{n \to \infty}{\longrightarrow} \infty$ telle que :

$$\forall \theta \in \Theta, \ \exists \Sigma_\theta \text{ symétrique semi-définie positive : } v_n(T_n - \theta) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta). \tag{3.5}$$

Σ_θ est une matrice $p \times p$ (p est la dimension de θ) ; c'est la matrice de covariance de la loi limite.

La normalité asymptotique implique la consistance (utiliser le lemme de Slutsky).

3.2 Estimateur par substitution

Théorème 5 (de l'application continue)**.** *Soit T un estimateur (fortement ou non) consistant de $\theta \in \Theta \subset \mathbb{R}^p$. Soit $\psi : \mathbb{R}^p \longrightarrow \mathbb{R}^d$ une fonction continue. Alors $\psi(T)$ est (fortement ou non) consistant pour $\psi(\theta)$.*

La propriété est également vraie pour la convergence en loi ; c'est alors une conséquence du théorème de Portmanteau. [Portmanteau veut dire *fourre tout* en anglais. Patrick Billingsley a écrit un ouvrage célèbre sur les convergences de variables aléatoires : *Convergence of Probability Measures*. Dans cet ouvrage, il attribue le théorème à Jean-Pierre Portmanteau, de l'université

de Felletin, petite commune française de la Creuse. L'article est daté de 1915 et intitulé *espoir pour l'ensemble vide*. Il s'agit d'un canular. Le théorème a en fait été prouvé par Alexandre Alexandrov vers 1940. Alexandrov était le directeur de thèse de Grigori Perelman, qui a obtenu la médaille Fields pour avoir démontré la conjecture de Poincaré. Il a refusé cette médaille, ainsi que le prix Wolf d'un million de dollars qui l'accompagnait.]

Théorème 6 (Méthode delta)**.** *Soit T_n un estimateur de θ tel que $v_n(T_n-\theta) \xrightarrow[n\to\infty]{\mathcal{L}} T \in \mathbb{R}^p$. Soit $\psi : \Theta \subset \mathbb{R}^p \longrightarrow \mathbb{R}^d$ une fonction différentiable de matrice jacobienne $J_\psi(\theta) \in \mathcal{M}_{d,p}(\mathbb{R})$ en θ. Alors*

$$v_n\left(\psi(T_n)-\psi(\theta)\right) \xrightarrow[n\to\infty]{\mathcal{L}} d\psi_\theta(T) = J_\psi(\theta)\times T, \tag{3.6}$$

et en particulier si T_n est asymptotiquement normal avec

$$v_n\left(T_n-\theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\theta \sim \mathcal{N}(0,\Sigma_\theta) \tag{3.7}$$

(Σ_θ est donc la matrice de variance-covariance de Z_θ), alors $\psi(T_n)$ est asymptotiquement normal et

$$v_n\left(\psi(T_n)-\psi(\theta)\right) \xrightarrow[n\to\infty]{\mathcal{L}} J_\psi(\theta)\times Z_\theta \sim \mathcal{N}\left(0, J_\psi(\theta)\times\Sigma_\theta\times J_\psi(\theta)^\top\right). \tag{3.8}$$

Remarques puis cas de la dimension 1 :

$d\psi_\theta(h)$ est la différentielle de ψ au point θ, calculée en h, on a $d\psi_\theta(h) = J_\psi(\theta)\times h$. Lorsque $d=1$, $J_\psi(\theta) = \nabla\psi_\theta^\top$ et donc la différentielle calculée en h est égale à $\nabla\psi_\theta^\top h$. Si de plus $p=1$, $\nabla\psi_\theta = \psi'(\theta)$ et $d\psi_\theta(h) = \psi'(\theta)h$. Dans ce dernier cas, on a simplement $J_\psi(\theta)\Sigma_\theta J_\psi(\theta)^\top = \psi'(\theta)^2\times\sigma_\theta^2$, où $\sigma_\theta^2 = \Sigma_\theta$ est la variance de la loi normale asymptotique.

Exemple :

Soit $(X_1,...,X_n)$ un n-échantillon d'une v.a. X de loi de Bernoulli de paramètre $\theta \in]0,1[$. Soit $\overline{X}_n$ la moyenne empirique de l'échantillon. D'après le théorème de la limite centrale,

$$\sqrt{n}\left(\overline{X}_n-\theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\theta \sim \mathcal{N}(0,\theta(1-\theta)) \tag{3.9}$$

Posons $\psi(\theta) = 2\arcsin\sqrt{\theta}$. $\psi(\overline{X}_n)$ est un estimateur de $\psi(\theta)$ et puisque ψ est dérivable, alors $\psi(\overline{X}_n)$ est asymptotiquement normal.

Comme

$$\psi'(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}},$$

Alors $J_\theta\Sigma_\theta J_\theta' = \psi'(\theta)^2\times\theta(1-\theta) = 1$ et l'on a donc

$$\sqrt{n}\left(2\arcsin\sqrt{\overline{X}_n} - 2\arcsin\sqrt{\theta}\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1).$$

Cette méthode, dite de "stabilisation de la variance", permet de construire un intervalle de confiance asymptotique pour θ (voir chapitre suivant).

Exemple :

Soit $(X_1,...,X_n)$ un n-échantillon d'une v.a. X de loi exponentielle de paramètre $\lambda>0$. Soit $\overline{X}_n$ la moyenne empirique de l'échantillon. On sait que

$$\mathbb{E}[X] = 1/\lambda \text{ et } \mathbb{V}(X) = 1/\lambda^2$$

D'après le théorème de la limite centrale, $\sqrt{n}\left(\overline{X}_n - 1/\lambda\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\lambda \sim \mathcal{N}(0,1/\lambda^2)$.

Posons $\psi(x) = 1/x$ qui est différentiable sur $]0,+\infty[$ avec $\psi'(x)^2 = 1/x^4$. En appliquant la méthode delta, on en déduit que

$$\sqrt{n}\left(\frac{1}{\overline{X}_n}-\lambda\right)\xrightarrow[n\to\infty]{\mathcal{L}} Z'_\lambda\sim\mathcal{N}\left(0,\frac{1}{\left(\frac{1}{\lambda}\right)^4}\frac{1}{\lambda^2}\right)=\mathcal{N}\left(0,\lambda^2\right),$$

qu'on peut réécrire comme :

$$\frac{\sqrt{n}}{\lambda}\left(\frac{1}{\overline{X}_n}-\lambda\right)\xrightarrow[n\to\infty]{\mathcal{L}} Z\sim\mathcal{N}(0,1).$$

Exemple :

Considérons un n-échantillon $(X_1,...,X_n)$ d'une v.a. X telle que $\mathbb{E}[X^4]<\infty$. On note $\alpha_i=\mathbb{E}[X^i]$ pour $i=1,\dots,4$ et $\alpha=(\alpha_1,\alpha_2)$. On s'intéresse à l'estimation de la variance $\alpha_2-\alpha_1^2$. On note également

$$\overline{X}_n=\widehat{m}_n^{(1)}=\tfrac{1}{n}\textstyle\sum_{i=1}^n X_i\quad,\quad \overline{X^2}_n=\widehat{m}_n^{(2)}=\tfrac{1}{n}\sum_{i=1}^n X_i^2\quad\text{et}\quad \hat{\sigma}_n^2=\tfrac{1}{n}\sum_{i=1}^n(X_i-\overline{X}_n)^2 \tag{3.10}$$

Soit $T_n=\left(\overline{X}_n,\overline{X^2}_n\right)^\top$. Soit $\psi(x,y)=y-x^2$. Alors $\hat{\sigma}_n^2=\psi(\overline{X}_n,\overline{X^2}_n)$. Le théorème de la limite centrale multidimensionnel s'applique et donne

$$\sqrt{n}(T_n-\alpha)=\sqrt{n}\left(\begin{pmatrix}\overline{X}_n\\ \overline{X^2}_n\end{pmatrix}-\begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\right)\xrightarrow[n\to\infty]{\mathcal{L}}\begin{pmatrix}T_1\\ T_2\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix},\begin{pmatrix}\alpha_2-\alpha_1^2 & \alpha_3-\alpha_1\alpha_2\\ \alpha_3-\alpha_1\alpha_2 & \alpha_4-\alpha_2^2\end{pmatrix}\right).$$

En effet, la matrice de variance-covariance de (X_1,X_1^2) est

$$\begin{aligned}\begin{pmatrix}\mathbb{V}(X_1) & Cov(X_1,X_1^2)\\ Cov(X_1,X_1^2) & \mathbb{V}(X_1^2)\end{pmatrix}&=\begin{pmatrix}\alpha_2-\alpha_1^2 & \mathbb{E}[X_1^3]-\mathbb{E}[X_1]\mathbb{E}[X_2]\\ \mathbb{E}[X_1^3]-\mathbb{E}[X_1]\mathbb{E}[X_2] & \alpha_4-\alpha_2^2\end{pmatrix}\\&=\begin{pmatrix}\alpha_2-\alpha_1^2 & \alpha_3-\alpha_1\alpha_2\\ \alpha_3-\alpha_1\alpha_2 & \alpha_4-\alpha_2^2\end{pmatrix}\end{aligned}$$

La fonction ψ est polynomiale donc différentiable et sa différentielle en (x,y) est $d\psi_{(x,y)}(h,k)=-2xh+k$. La méthode delta appliquée à ψ donne alors

$$\sqrt{n}\left(\hat{\sigma}_n^2-(\alpha_2-\alpha_1^2)\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(0,-4\alpha_1^4-\alpha_2^2+8\alpha_1^2\alpha_2-4\alpha_1\alpha_3+\alpha_4\right) \tag{3.11}$$

Cette loi normale est la loi de la variable aléatoire $-2\alpha_1T_1+T_2=\begin{pmatrix}-2\alpha_1 & 1\end{pmatrix}\begin{pmatrix}T_1\\ T_2\end{pmatrix}$ où (T_1,T_2) est le vecteur limite de $\sqrt{n}(T_n-\alpha)$.

Lorsque les variables sont centrées, c'est à dire lorsque $\alpha_1=0$, le résultat devient simplement

$$\sqrt{n}\left(\hat{\sigma}_n^2-\alpha_2\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(0,\alpha_4-\alpha_2^2\right). \tag{3.12}$$

Exemple :

Soit $(X_1,Y_1),...,(X_n,Y_n)$ un n-échantillon d'un vecteur gaussien (X,Y) dont le coefficient de corrélation linéaire est noté ρ. Le coefficient de corrélation empirique est par définition égal à

$$\widehat{\rho}_n=\frac{\sum_{i=1}^n\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\left[\sum_{i=1}^n\left(X_i-\overline{X}\right)^2\times\sum_{i=1}^n\left(Y_i-\overline{Y}\right)^2\right]^{1/2}} \tag{3.13}$$

Le théorème de la limite centrale et la méthode delta montrent que $\sqrt{n}(\widehat{\rho}_n-\rho)$ est asymptotiquement normale. On peut montrer (à faire en exercice, les calculs sont difficiles, s'aider de Wikipedia ici et ici) que

$$\sqrt{n}(\widehat{\rho}_n-\rho)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}\left(0,(1-\rho^2)^2\right). \tag{3.14}$$

La méthode de stabilisation de la variance amène à chercher une fonction différentiable ψ dont la dérivée vaut $(1-\rho^2)^{-1}$. On pose donc naturellement $\psi(\rho)=\operatorname{arctanh}\rho$ et une nouvelle application de la méthode delta montre que

$$\sqrt{n}(\psi(\widehat{\rho}_n)-\psi(\rho))\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,1). \tag{3.15}$$

3.3 Estimateur des moments

On rappelle qu'un estimateur par la méthode des moments est de la forme $\widehat{\theta}_n = f\left(\widehat{m}_n^{(1)}, \dots, \widehat{m}_n^{(k)}\right) = f\left(\widehat{M}_n^{(k)}\right)$ avec $k \geq 1$, $\widehat{M}_n^{(k)} = \begin{pmatrix} \widehat{m}_n^{(1)} \\ \vdots \\ \widehat{m}_n^{(k)} \end{pmatrix}$, et f une fonction telle que $\theta = f\left(M^{(k)}\right)$, où $M^{(k)} = \begin{pmatrix} m^{(1)} \\ \vdots \\ m^{(k)} \end{pmatrix}$. Donc c'est un estimateur par substitution substituant des moyennes empiriques, on peut donc en déduire des propriétés asymptotiques avec la LGN, le LAC, le TLC (éventuellement multidimensionnel), et la méthode delta.

On présente ensuite une trame générale de raisonnement **à appliquer au cas par cas dans les applications.**

Si la loi commune de l'échantillon admet un moment d'ordre k, la LGN appliquée k fois donne $\widehat{m}_n^{(i)} \xrightarrow[n\to\infty]{\mathbb{P}} m^{(i)}$ pour tout $i \in [\![1, k]\!]$, par propriétés de la convergence en probabilités on a ensuite $\widehat{M}_n^{(k)} \xrightarrow[n\to\infty]{\mathbb{P}} M^{(k)}$. Si f est continue en θ on a alors par le LAC $\widehat{\theta}_n = f\left(\widehat{M}_n^{(k)}\right) \xrightarrow[n\to\infty]{\mathbb{P}} f\left(M^{(k)}\right) = \theta$. Si de plus la loi commune de l'échantillon admet un moment d'ordre $2k$, le TLC multidimensionnel donne

$$\sqrt{n}\left(\widehat{M}_n^{(k)} - M^{(k)}\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\theta \sim \mathcal{N}\left(0, \Sigma_\theta\right),$$

avec $\Sigma_\theta = \left(\mathbb{C}\mathrm{ov}(X_1^i, X_1^j)\right)_{1\leq i,j\leq k}$ qui est la matrice de variance-covariance de $\begin{pmatrix} X_1 \\ X_1^2 \\ \vdots \\ X_1^k \end{pmatrix}$. Enfin, si f est différentiable en θ, alors la delta-méthode donne la normalité asymptotique de $\widehat{\theta}_n$:

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathrm{d}f_{M^{(k)}}\left(Z_\theta\right) = J_f\left(M^{(k)}\right) Z_\theta \sim \mathcal{N}\left(0, J_f\left(M^{(k)}\right) \Sigma_\theta \left(J_f\left(M^{(k)}\right)\right)^\top\right).$$

Exemple :

Soit $(X_1, ..., X_n)$ un n-échantillon de $X \sim B(\alpha, \beta)$, dont la densité est

$$g(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}_{[0,1]}(x)$$

avec $\alpha, \beta \in [0, 1]$. L'estimateur des moments de (α, β) est calculé en trouvant la solution de

$$\begin{cases} m^{(1)}(\alpha,\beta) = \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \\ m^{(2)}(\alpha,\beta) = \mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \end{cases}$$

puis en substituant $(m^{(1)}, m^{(2)})$ par $(\widehat{m}_n^{(1)}, \widehat{m}_n^{(2)}) = (\overline{X}_n, \overline{X^2}_n)$.

La fonction

$$\psi(x, y) = \left(\frac{x}{x+y}, \frac{x(x+1)}{(x+y)(x+y+1)}\right)$$

est de classe C^1 et inversible sur $]0, 1[^2$. En inversant cette fonction, on trouve l'expression des estimateurs en fonction de $\overline{X}$ et $\overline{X^2}$:

$$\begin{cases} \widehat{\alpha} = \dfrac{\overline{X}_n(\overline{X^2}_n - \overline{X}_n)}{\overline{X}_n^2 - \overline{X^2}_n} \\ \widehat{\beta} = \dfrac{(1-\overline{X}_n)(\overline{X^2}_n - \overline{X}_n)}{\overline{X}_n^2 - \overline{X^2}_n} \end{cases}$$

et ces estimateurs sont consistants et asymptotiquement normaux (à faire en exercice, ψ n'est pas facile à inverser quand on ne sait pas ce qu'on doit trouver, mais là c'est donné donc il suffit de vérifier que c'est bien l'inverse. Sinon astuce : remplacer directement le $\alpha/(\alpha+\beta)$ par $m^{(1)}$ (le système devient linéaire) et éventuellement changer d'inconnue en posant $\gamma = \alpha + \beta$).

3.4 Estimateur du maximum de vraisemblance

C'est Fisher qui a le premier étudié, en 1921 et 1925, l'estimateur du maximum de vraisemblance. En 1946, Cramer a donné une première preuve de la consistance de cet estimateur et de sa normalité asymptotique, sous certaines conditions de régularité. Wald a démontré la consistance forte en 1949, sous des hypothèses plus générales. Rappelons que $\nabla_\theta \ell(\theta; X)$ s'appelle le score.

Il existe une foultitude de conditions de régularité différentes selon les ouvrages, qui conduisent à la consistance forte ou faible de l'e.m.v. Voici un exemple d'énoncé de consistance de l'EMV.

On note $L(\theta; X)$ la vraisemblance d'une observation, $\ell(\theta; X)$ son logarithme, dans ce cas X est donc une variable aléatoire réelle, et $L_n(\theta; X)$ la vraisemblance de l'échantillon de taille n, dans ce cas X est un vecteur aléatoire de taille n constitué de variables i.i.d. On se place ici dans le cas unidimensionnel mais le cas multidimensionnel se traite de manière identique.

Théorème 7 (Consistance de l'e.m.v.)**.** *On considère le modèle d'échantillonnage $X_1, \dots, X_n$ de v.a.i.i.d. $\sim P_{\theta^*}$. Sous les hypothèses de régularité ci-dessous,*

- *Le modèle est identifiable.*
- *Θ est un compact de $\mathbb{R}^p$.*
- *Pour tout $x \in \mathbb{X}$, la fonction $\theta \in \Theta \mapsto L(\theta, x)$ est continue.*
- *La fonction $x \mapsto \sup_{s\in\Theta} |\ell(s; x)|$ est P_θ-intégrable pour tout $\theta \in \Theta$.*

Alors l'estimateur du maximum de vraisemblance est consistant.

À noter que l'hypothèse que Θ est compact est très restrictive et rarement rencontrée dans nos exemples.

La normalité asymptotique de l'e.m.v. requiert des conditions supplémentaires. Avant cela, il est nécessaire d'introduire la quantité dite *Information de Fisher*.

Définition 19. *Supposons que la fonction $\theta \mapsto \ell(\theta; x)$ est différentiable pour presque tout x. Alors la fonction $I(\cdot)$ définie sur Θ :*

$$I(\theta) = \int \ell'(\theta; x)^2 P_\theta(\mathrm{d}x) = \int \ell'(\theta; x)^2 L(\theta; x)\mu(\mathrm{d}x) = \mathbb{E}_\theta[(\ell'(\theta; X))^2]$$

est appelée Information de Fisher *associée à une observation. Dans le cas multidimensionnel l'Information de Fisher devient une matrice $\mathbb{E}_\theta[\nabla_\theta \ell(\theta; X) \nabla_\theta \ell(\theta; X)^\top]$.*

Afin d'aller plus loin, il est nécessaire de fixer les conditions supplémentaires. Celles-ci portent sur la régularité du modèle.

Définition 20. *Un modèle paramétrique est régulier si, et seulement si,*

- *Il est dominé et homogène.*
- *Pour μ-presque tout $x \in \mathbb{X}$, les fonctions $\theta \mapsto L(\theta; x)$ et $\theta \mapsto \ell(\theta; x)$ sont deux fois continûment dérivables (C^2) sur $\mathring{\Theta}$.*
- *Pour tout $\theta^* \in \mathring{\Theta}$ il existe un voisinage U de θ^* dans $\mathring{\Theta}$ et une fonction borélienne $\Lambda(\cdot)$ tels que $|\ell''(s; x)| \le \Lambda(x), |\ell'(s; x)| \le \Lambda(x), \left|\ell'(s; x)^2\right| \le \Lambda(x)$, pour tout $s \in U$ et μ-presque tout $x \in \mathbb{X}$, et*

$$\int \Lambda(x) \sup_{\theta\in U} L(\theta; x)\mathrm{d}\mu(x) < \infty.$$

- *L'information de Fisher $I(\theta)$ est strictement positive pour tout $\theta \in \mathring{\Theta}$.*

Le troisième point assure l'existence de l'Information de Fisher et aussi la possibilité de dériver jusqu'à deux fois en θ sous le signe intégrale (et de permuter intégrale et dérivée), par conséquence du théorème de convergence dominée. Il se généralise en multidimensionnel en utilisant toutes les dérivées partielles. Le quatrième point se généralise en demandant que l'information de Fisher soit définie positive.

Les modèles Gaussien ou de Poisson sont réguliers, mais pas le modèle uniforme sur $[0, \theta]$ (car il n'est pas homogène). Habituellement, on ne vérifie pas les conditions techniques et cela ne sera pas demandé. Elles sont vérifiées dans tous les cas d'exercice sauf éventuellement la

question du support, qui est primordiale et doit être vérifiée, elle. Enfin, il faut quand même citer la régularité du modèle quand elle est utilisée, par exemple quand on utilise les théorèmes suivants.

Théorème 8. *Si le modèle est régulier, l'information de Fisher du modèle est la variance du score (car dans ce cas le score est centré) et est aussi l'espérance de la dérivée seconde de la log-vraisemblance :*

$$I(\theta) = -\mathbb{E}_\theta[\ell''(\theta; X)]. \tag{3.16}$$

Dans le cas multidimensionnel, on utilise la hessienne et alors $I(\theta) = -\mathbb{E}_\theta[H_\ell(\theta)]$.

Démonstration.

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\ell(\theta;X)\right] &= \int \frac{\partial}{\partial\theta_i}\ell(\theta;x)L(\theta;x)\mathrm{d}\mu(x)\\
&= \int \frac{\partial}{\partial\theta_i}L(\theta;x)\frac{1}{L(\theta;x)}L(\theta;x)\mathrm{d}\mu(x)\\
&= \int \frac{\partial}{\partial\theta_i}L(\theta;x)\mathrm{d}\mu(x)\\
&= \frac{\partial}{\partial\theta_i}\int L(\theta;x)\mathrm{d}\mu(x)\\
&= \frac{\partial}{\partial\theta_i}1 = 0.
\end{aligned}$$

donc le score est bien centré et l'Information de Fisher est bien sa matrice de variance-covariance.

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\ell(\theta;X)\frac{\partial}{\partial\theta_j}\ell(\theta;X)\right] &= \int \frac{\partial}{\partial\theta_i}\ell(\theta;x)\frac{\partial}{\partial\theta_j}\ell(\theta;x)L(\theta;x)\mathrm{d}\mu(x)\\
&= \int \frac{\partial}{\partial\theta_i}\ell(\theta;x)\frac{\partial}{\partial\theta_j}L(\theta;x)\frac{1}{L(\theta;x)}L(\theta;x)\mathrm{d}\mu(x)\\
&= \int \frac{\partial}{\partial\theta_i}\ell(\theta;x)\frac{\partial}{\partial\theta_j}L(\theta;x)\mathrm{d}\mu(x)\\
&= \int \frac{\partial}{\partial\theta_j}\left(\frac{\partial}{\partial\theta_i}\ell(\cdot;x)L(\cdot;x)\right)(\theta)\mathrm{d}\mu(x) - \int\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta;x)\right)L(\theta;x)\mathrm{d}\mu(x)\\
&= \int \frac{\partial}{\partial\theta_j}\left(\frac{\partial}{\partial\theta_i}L(\theta;x)\right)\mathrm{d}\mu(x) - \int\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta;x)\right)L(\theta;x)\mathrm{d}\mu(x)\\
&= \frac{\partial}{\partial\theta_j}\frac{\partial}{\partial\theta_i}\int L(\theta;x)\mathrm{d}\mu(x) - \int\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta;x)\right)L(\theta;x)\mathrm{d}\mu(x)\\
&= 0 - \int\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta;x)\right)L(\theta;x)\mathrm{d}\mu(x)
\end{aligned}$$

□

Cela peut se voir comme la courbure de la vraisemblance limite. C'est l'information de Fisher en θ^* qui apparaît après ; si cette quantité est grande, cela veut dire que la vraisemblance limite est "piquée" en ce point, il est donc plus facile d'estimer le paramètre. On peut enfin énoncer la normalité asymptotique :

Théorème 9. *On considère un modèle d'échantillonage issu d'un modèle régulier. La suite d'estimateurs du maximum de vraisemblance est notée $\hat\theta_n$. On suppose que pour tout $\theta^* \in \Theta$, $\hat\theta_n$ est consistant. Alors, pour tout $\theta^* \in \mathring{\Theta}$, on a :*

$$\sqrt{n}\left(\hat\theta_n - \theta^*\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, I^{-1}(\theta^*)\right).$$

(Hors programme :) L'estimateur du maximum de vraisemblance est dit "asymptotiquement efficace" car on peut montrer qu'aucun estimateur sans biais ne peux avoir une variance plus faible (c'est la borne de Fréchet-Darmois-Cramer-Rao). Néanmoins, ce résultat est peu utile.

Chapitre 4

Estimation par intervalle de confiance

4.1 Intervalle et région de confiance : définitions

Définition 21. *Soit $\alpha \in]0,1[$. Une région de confiance au niveau de confiance $1-\alpha$ pour l'estimation de $\theta \in \Theta$ notée $RC_{1-\alpha}(\cdot,\theta)$ ou simplement $RC_{1-\alpha}(\theta)$ (sans la dépendance en X) est une fonction $: \mathbb{X} \to \mathcal{P}(\Theta')$ avec $\Theta \subseteq \Theta'$, qui ne dépend pas de θ, telle que $\{\theta \in RC_{1-\alpha}(X,\theta)\}$ est mesurable pour tout $\theta \in \Theta$ et tout $X \sim P_\theta$ et telle que :*

$$\forall \theta \in \Theta, \mathbb{P}_\theta\left(\theta \in RC_{1-\alpha}(\theta)\right) \geq 1-\alpha. \tag{4.1}$$

Elle est exacte si il y a égalité dans (4.1)*. Elle est plutôt appelée intervalle de confiance et notée $IC_{1-\alpha}(X,\theta)$ si $\theta \subseteq \mathbb{R}$ et $RC_{1-\alpha}(x,\theta)$ est un intervalle pour tout $x \in \mathbb{X}$.*

La notion se généralise bien entendu à l'estimation de n'importe quelle fonction de θ, $h(\theta)$.

Attention à la notation trompeuse, $RC_{1-\alpha}(\theta)$ et $IC_{1-\alpha}(\theta)$ ne dépendent pas de θ qui est inconnu mais de X, le θ est là pour indiquer que c'est un intervalle de confiance pour θ.

Ne pas non plus oublier que le $\mathbb{P}_\theta$ dans (4.1) est abusif et devrait juste être un $\mathbb{P}$. Le $_\theta$ est là pour éviter d'écrire en fait "$\forall X \sim P_\theta, \mathbb{P}\left(\theta \in RC_{1-\alpha}(\theta)\right) \geq 1-\alpha$".

Les deux critères de qualité d'un intervalle de confiance (sa longueur et son niveau de confiance) s'opposent et il faut donc réaliser un compromis entre les deux. On cherche généralement un intervalle de longueur la plus petite possible, pour un niveau de risque α donné, ce qui conduit à désirer une région exacte. Pour comparer deux intervalles de confiance au même niveau de confiance, on comparera leur longueur et le meilleur sera le plus court.

4.2 Retour sur la fonction quantile

La fonction quantile est un outil qui va être très utilisé pour la construction d'intervalles de confiance mais aussi de tests statistiques.

Étant donné la fonction de répartition F d'une v.a. réelle X de loi $\mathcal{L}$, on rappelle que la fonction quantile associée, notée Q_F, Q_X, $Q_\mathcal{L}$, ou Q s'il n'y a pas d'ambiguïté, est définie par :

$$Q : \left\{ \begin{array}{ccc}]0,1[& \longrightarrow & \mathbb{R} \\ \alpha & \longmapsto & \inf\{x \in \mathbb{R} : F(x) \geq \alpha\} \end{array} \right. \tag{4.2}$$

$Q(\alpha)$ est souvent aussi noté q_α^F, q_α^X, $q_\alpha^\mathcal{L}$ ou q_α s'il n'y a pas d'ambiguïté.

Proposition 8.

1. *Q est toujours bien définie.*
2. *Q est en fait un minimum, c'est-à-dire que $F(q_\alpha) \geq \alpha$ pour tout $\alpha \in]0,1[$.*
3. *On a aussi, pour tout $\alpha \in]0,1[$, $F^-(q_\alpha) \leq \alpha$.*
4. *Si F est continue en q_α, alors $F(q_\alpha) = \alpha$.*
5. *Si F est bijective alors $F = Q^{-1}$.*

Démonstration.

1. Soit $\alpha \in]0,1[$, en particulier $\alpha < 1$. $\lim_{x\to\infty} F(x) = 1$ donc $\exists A : F(A) > \alpha$ et $\{x \in \mathbb{R} : F(x) \geq \alpha\}$ est non-vide. Mais on a aussi $\alpha > 0$ et $\lim_{x\to-\infty} F(x) = 0$ donc $\exists B : \forall b \leq B, F(b) < \alpha$, donc $\{x \in \mathbb{R} : F(x) \geq \alpha\}$ est minorée par B. Toute partie de $\mathbb{R}$ non-vide et minorée possède une borne inférieure, q_α existe bien et Q est bien définie.
2. Soit $\alpha \in]0,1[$, soit $\varepsilon > 0$. Par propriété des bornes inférieures il existe $y \in]q_\alpha, q_\alpha + \varepsilon[$ tel que $y \in \{x \in \mathbb{R} : F(x) \geq \alpha\}$, soit $F(y) \geq \alpha$. Par croissance, $F(q_\alpha + \varepsilon) \geq \alpha$. On fait tendre ε vers 0, par continuité à droite il vient $F(q_\alpha) \geq \alpha$.
3. Soit $\alpha \in]0,1[$, soit $\varepsilon > 0$. Par l'absurde si $F(q_\alpha - \varepsilon) \geq \alpha$ alors $q_\alpha - \varepsilon \in \{x \in \mathbb{R} : F(x) \geq \alpha\}$ dont q_α est le minimum : $q_\alpha \leq q_\alpha - \varepsilon$ et on a une contradiction, donc $F(q_\alpha - \varepsilon) < \alpha$. On fait tendre ε vers 0, il vient $F^-(q_\alpha) \leq \alpha$.
4. On combine juste les deux points précédents.
5. Si F est bijective alors par injectivité elle est strictement croissante et donc $F(x) \geq \alpha$ si et seulement si $x \geq F^{-1}(\alpha)$, donc $Q(\alpha) = \inf\left\{x \in \mathbb{R} : x \geq F^{-1}(\alpha)\right\} = \inf\left[F^{-1}(\alpha), \infty\right[= F^{-1}(\alpha)$.

□

4.3 Méthodes pour déterminer une région de confiance

4.3.1 Fonction pivotale

Définition 22. *Un pivot $Z = Z(X, \theta)$ est une fonction de la variable d'observation X et du paramètre θ telle que la loi de Z est libre en θ, c'est à dire que sa loi ne dépend pas de θ.*

On dit que θ est un paramètre de position pour T si la loi de $T - \theta$ ne dépend pas de θ, c'est à dire si $T - \theta$ est un pivot.

On dit que θ est un paramètre d'échelle pour T si la loi de T/θ ne dépend pas de θ, c'est à dire si T/θ est un pivot.

Un pivot est donc différent d'une statistique : une statistique ne dépend que de X et sa loi peut dépendre de θ, tandis que le pivot peut dépendre des deux mais sa loi ne doit pas dépendre de θ. Pour ajouter à la confusion, un pivot est parfois appelé "statistique pivotale".

Une fois un pivot déterminé, on peut déterminer un évènement B (qui ne dépend pas non plus de θ) tel que $\mathbb{P}_\theta(Z \in B) \geq 1 - \alpha$. La région de confiance est alors de la forme $RC_{1-\alpha}(\theta) = \{\tilde{\theta} \in \Theta : Z(X, \tilde{\theta}) \in B\}$ et il s'agit ensuite de résoudre l'inéquation $Z(X, \tilde{\theta}) \in B$ en $\tilde{\theta}$. Cette méthode en 3 étapes (déterminer Z, puis B, puis $RC_{1-\alpha}(\theta)$) s'appelle la méthode du pivot.

Démonstration. On montre que $\{\tilde{\theta} \in \Theta : Z(X, \tilde{\theta}) \in B\}$ est bien une région de confiance.

$$\begin{aligned}\mathbb{P}_\theta\left(\theta \in RC_{1-\alpha}(\theta)\right) &= \mathbb{P}_\theta\left(\theta \in \{\tilde{\theta} \in \Theta : Z(X, \tilde{\theta}) \in B\}\right) \\ &= \mathbb{P}_\theta\left(Z(X, \theta) \in B\right) \\ &\geq 1 - \alpha.\end{aligned}$$

□

Un cas que l'on va rencontrer très fréquemment est celui où Z est une v.a.r. Notons $\mathcal{L}$ sa loi (qui pour rappel ne dépend pas de θ). Dans ce cas, les 3 événements suivants conviennent.

Proposition 9.

1. $B_1 = \left[q^{\mathcal{L}}_{\frac{\alpha}{2}}, q^{\mathcal{L}}_{1-\frac{\alpha}{2}}\right]$.
2. $B_2 = \left[q^{\mathcal{L}}_{\alpha}, \infty\right[$.
3. $B_3 = \left]-\infty, q^{\mathcal{L}}_{1-\alpha}\right]$.

Démonstration. Soit F la c.d.f. de $\mathcal{L}$.

$$\begin{aligned}\mathbb{P}(Z \in B_1) &= \mathbb{P}\left(q^{\mathcal{L}}_{\frac{\alpha}{2}} \leq Z \leq q^{\mathcal{L}}_{1-\frac{\alpha}{2}}\right) \\ &= F\left(q^{\mathcal{L}}_{1-\frac{\alpha}{2}}\right) - F^{-}\left(q^{\mathcal{L}}_{\frac{\alpha}{2}}\right) \\ &\geq 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1-\alpha.\end{aligned}$$

Les autres cas se traitent encore plus simplement. □

Souvent on aura $\Theta \subseteq \mathbb{R}$ et la résolution de l'inéquation donnera un intervalle de confiance. L'intervalle sera dit bilatéral s'il est construit à partir de B_1, et unilatéral s'il est construit à partir de B_2 ou B_3. De plus si la loi $\mathcal{L}$ est continue, alors les régions ainsi construites sont exactes.

La rédaction typique à adopter est la suivante, sachant que le pivot se construit généralement à partir d'un estimateur $\hat{\theta}$ de θ : "Soit $Z = Z(\hat{\theta}, \theta)$, sa loi est $\mathcal{L}$ quel que soit $\theta \in \Theta$. On a donc $\mathbb{P}\left(q^{\mathcal{L}}_{\frac{\alpha}{2}} \leq Z(\hat{\theta}, \theta) \leq q^{\mathcal{L}}_{1-\frac{\alpha}{2}}\right) \geq 1-\alpha$, donc $\mathbb{P}\left(\hat{a} \leq \theta \leq \hat{b}\right) \geq 1-\alpha$ (résolution de l'inéquation, $\hat{a}$ et $\hat{b}$ dépendant de $\hat{\theta}$), on en déduit donc qu'un intervalle de confiance au niveau $1-\alpha$ pour θ est $IC_{1-\alpha}(\theta) = [\hat{a}, \hat{b}]$."

Exemple :

Soit $(X_1, ..., X_n)$ un n-échantillon de loi $\mathcal{N}(m, \sigma^2)$.

• Supposons m inconnu et σ^2 connu et intéressons-nous à l'estimation de m.

Un estimateur de m est donné par $\overline{X}_n \sim \mathcal{N}(m, \sigma^2/n)$. On peut alors choisir comme pivot

$$Z_n = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Soit $\alpha \in]0,1[$, on a $\mathbb{P}\left(q^*_{\frac{\alpha}{2}} \leq Z_n \leq q^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$. De plus, par symétrie de la loi $\mathcal{N}(0,1)$, $q^*_{\frac{\alpha}{2}} = -q^*_{1-\frac{\alpha}{2}}$. Donc

$$\begin{aligned}1-\alpha &= \mathbb{P}\left(-q^*_{1-\frac{\alpha}{2}} \leq Z_n \leq q^*_{1-\frac{\alpha}{2}}\right) \\ &= \mathbb{P}\left(-q^*_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}_n - m}{\sigma/\sqrt{n}} \leq q^*_{1-\frac{\alpha}{2}}\right) \\ &= \mathbb{P}\left(\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^*_{1-\frac{\alpha}{2}} \leq m \leq \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^*_{1-\frac{\alpha}{2}}\right)\end{aligned}$$

Un intervalle de confiance est alors

$$IC^{(1)}_{1-\alpha}(m) = \left[\overline{X}_n - \frac{\sigma}{\sqrt{n}} q^*_{1-\frac{\alpha}{2}}, \overline{X}_n + \frac{\sigma}{\sqrt{n}} q^*_{1-\frac{\alpha}{2}}\right]$$

• Supposons m et σ^2 inconnus et intéressons-nous à l'estimation de m.

La v.a. $Z_n = (\overline{X}_n - \theta)/(\sigma/\sqrt{n})$ définie précédemment suit une loi $\mathcal{N}(0,1)$ mais elle fait intervenir σ^2 qui est inconnu, donc l'IC construit précédemment n'est cette fois pas calculable et ne convient pas. Posons

$$Y_n = \frac{(n-1)S_n^2}{\sigma^2}$$

où S_n^2 est la variance empirique non biaisée. D'après le théorème de Cochran, Y_n suit une loi $\chi^2(n-1)$ et $S_n^2 \perp \overline{X}_n$. Nous en déduisons que

$$T_n = \frac{Z_n}{\sqrt{Y_n/(n-1)}} = \frac{\overline{X}_n - m}{\sqrt{S_n^2/n}}$$

suit une loi de Student à $n-1$ degrés de libertés, ne dépend que de m (mais sa loi n'en dépend pas, elle est libre de m) et peut donc être choisie comme pivot. Par symétrie de la loi de Student,

le quantile $q_{1-\frac{\alpha}{2}}^{\mathcal{T}(n-1)}$ d'ordre $1-\alpha/2$ de cette loi de Student est tel que $-q_{1-\frac{\alpha}{2}}^{\mathcal{T}(n-1)}$ est le quantile d'ordre $\alpha/2$. Un intervalle de confiance est donc

$$IC_{1-\alpha}^{(2)}(m) = \left[\overline{X}_n - \sqrt{\frac{S_n^2}{n}} q_{1-\frac{\alpha}{2}}^{\mathcal{T}(n-1)}, \overline{X}_n + \sqrt{\frac{S_n^2}{n}} q_{1-\frac{\alpha}{2}}^{\mathcal{T}(n-1)}\right]$$

• Supposons m et σ^2 inconnus et intéressons-nous à l'estimation de σ^2.
Un intervalle de confiance est alors, d'après ce qui précède sur Y_n,

$$IC_{1-\alpha}^{(3)}(\sigma^2) = \left[\frac{(n-1)S_n^2}{q_{1-\frac{\alpha}{2}}^{\chi^2(n-1)}}, \frac{(n-1)S_n^2}{q_{\frac{\alpha}{2}}^{\chi^2(n-1)}}\right].$$

• Supposons m et σ^2 inconnus et intéressons-nous à l'estimation de (m, σ^2).

C'est une région de confiance incluse dans $\mathbb{R}^2$ que nous allons déterminer. Posons

$$C_n = (Z_n, Y_n) \sim \mathcal{N}(0,1) \otimes \chi^2(n-1)$$

Les deux composantes de Z étant indépendantes,

$$\begin{aligned}\mathbb{P}\left(C_n \in \left[-q_{1-\frac{\beta}{2}}^*, q_{1-\frac{\beta}{2}}^*\right] \times \left[q_{\frac{\beta}{2}}^{\chi^2(n-1)}, q_{1-\frac{\beta}{2}}^{\chi^2(n-1)}\right]\right) &= \mathbb{P}\left(Z_n \in \left[-q_{1-\frac{\beta}{2}}^*, q_{1-\frac{\beta}{2}}^*\right]\right) \mathbb{P}\left(Y_n \in \left[q_{\frac{\beta}{2}}^{\chi^2(n-1)}, q_{1-\frac{\beta}{2}}^{\chi^2(n-1)}\right]\right) \\ &= (1-\beta)^2 \\ &= 1-\alpha\end{aligned}$$

en posant $\beta = 1 - \sqrt{1-\alpha}$. La région de confiance sera donc de la forme

$$RC_{1-\alpha}((m,\sigma^2)) = \left\{(\tilde{m}, \tilde{\sigma}^2) : \tilde{\sigma}^2 \in IC_{1-\beta}^{(3)}(\sigma^2), \overline{X}_n - \frac{\tilde{\sigma}}{\sqrt{n}} q_{1-\frac{\beta}{2}}^* \leq \tilde{m} \leq \overline{X}_n + \frac{\tilde{\sigma}}{\sqrt{n}} q_{1-\frac{\beta}{2}}^*\right\}.$$

4.3.2 Méthode de Bonferroni

On suppose que $\Theta \subseteq \mathbb{R}^d$, et qu'on sait construire, pour tout niveau de confiance $1-\alpha$, $RC_{1-\alpha}(\theta_k)$ pour $1 \leq k \leq d$. C'est-à-dire que pour tout $k \leq d$, tout $\theta \in \Theta$, $\mathbb{P}_\theta\left(\theta_k \in RC_{1-\alpha}(\theta_k)\right) \geq 1-\alpha$. Alors une région de confiance au niveau de confiance $1-\alpha$ pour $\theta = (\theta_1, \ldots, \theta_d)$ est

$$\bigtimes_{k=1}^{d} RC_{1-\frac{\alpha}{d}}(\theta_k).$$

Démonstration.

$$\begin{aligned}\mathbb{P}_\theta\left(\theta \in \bigtimes_{k=1}^{d} RC_{1-\frac{\alpha}{d}}(\theta_k)\right) &= 1 - \mathbb{P}_\theta\left(\theta \notin \bigtimes_{k=1}^{d} RC_{1-\frac{\alpha}{d}}(\theta_k)\right) \\ &= 1 - \mathbb{P}_\theta\left(\bigcup_{k=1}^{d} \{\theta_k \notin RC_{1-\frac{\alpha}{d}}(\theta_k)\}\right) \\ &\geq 1 - \sum_{k=1}^{d} \mathbb{P}_\theta\left(\theta_k \notin RC_{1-\frac{\alpha}{d}}(\theta_k)\right) \\ &\geq 1 - \sum_{k=1}^{d} \frac{\alpha}{d} \\ &\geq 1-\alpha.\end{aligned}$$

□

À noter qu'à cause de l'inégalité dans la borne d'union, la méthode de Bonferroni ne donne en général pas de région de confiance exacte.

Exemple :

Soit $(X_1, ..., X_n)$ un n-échantillon de loi $\mathcal{N}(m, \sigma^2)$.

• Supposons m et σ^2 inconnus et intéressons-nous à l'estimation de (m, σ^2).

Par la méthode de Bonferroni, une région de confiance est donnée par

$$IC^{(2)}_{1-\frac{\alpha}{2}}(m) \times IC^{(3)}_{1-\frac{\alpha}{2}}(\sigma^2) = \left[\overline{X}_n - \sqrt{\frac{S_n^2}{n}} q_{1-\frac{\alpha}{4}}^{\mathcal{T}(n-1)}, \overline{X}_n + \sqrt{\frac{S_n^2}{n}} q_{1-\frac{\alpha}{4}}^{\mathcal{T}(n-1)}\right] \times \left[\frac{(n-1)S_n^2}{q_{1-\frac{\alpha}{4}}^{\chi^2(n-1)}}, \frac{(n-1)S_n^2}{q_{\frac{\alpha}{4}}^{\chi^2(n-1)}}\right]$$

4.3.3 Utilisation des inégalités de Tchebychev et de Hoeffding

Lorsque la v.a. est discrète, que la loi n'est pas suffisamment connue, ou qu'on ne trouve pas de pivot, il se peut qu'on ne puisse pas construire d'intervalle de confiance exact. Mais on peut parfois utiliser des inégalités de concentration pour construire tout de même des intervalles de confiance non exacts.

Si on a un résultat de concentration du type $\mathbb{P}_\theta\left(|\hat{\theta}_n - \theta| \geq \varepsilon\right) \leq f(\varepsilon)$, on peut inverser la relation du terme de droite en posant $f(\varepsilon) = \alpha$, on en tire une relation du type $\mathbb{P}_\theta\left(|\hat{\theta}_n - \theta| \geq \varepsilon(\alpha)\right) \leq \alpha$ et alors on peut poser $IC_{1-\alpha}(\theta) = \left]\hat{\theta}_n - \varepsilon(\alpha), \hat{\theta}_n + \varepsilon(\alpha)\right[$. $\varepsilon(\alpha)$ peut dépendre aussi de $(X_1, \ldots, X_n)$ mais pas de θ.

Théorème 10 (Inégalité de Bienaymé-Tchebychev)**.** *Soit X une v.a.r. qui appartient à L^2. Alors pour tout $\varepsilon > 0$,*

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq \varepsilon\right) \leq \frac{\mathbb{V}(X)}{\varepsilon^2} \tag{4.3}$$

Théorème 11 (Inégalité de Hoeffding)**.** *Soient $X_1, ..., X_n$ n v.a.i.i.d. telles que $a \leq X_1 \leq b$ $\mathbb{P}$-p.s. et $S_n = \sum_{i=1}^n X_i$. Alors pour tout $t > 0$,*

$$\mathbb{P}\left(|S_n - \mathbb{E}[S_n]| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right) \tag{4.4}$$

En revenant à $\overline{X}_n$ on a :

$$\begin{aligned}\mathbb{P}(\left|\overline{X}_n - \mathbb{E}[X]\right| \geq t) &= \mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq nt) \\ &\leq 2\exp\left(-\frac{2n^2t^2}{n(b-a)^2}\right) \\ &\leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).\end{aligned}$$

Exemple :

Considérons un n-échantillon $X_1, ..., X_n$ dont la loi est à support dans un intervalle $[a, b]$ fixé (en particulier il admet un moment de n'importe quel ordre car il est borné). Cherchons un intervalle de confiance pour la moyenne commune m des X_i.

L'inégalité de Tchebychev donne comme intervalle de confiance (non-exact) de niveau $1 - \alpha$

$$I_1 = \left]\overline{X}_n - \frac{b-a}{\sqrt{n\alpha}} \ ; \ \overline{X}_n + \frac{b-a}{\sqrt{n\alpha}}\right[.$$

En effet $|X_1 - \mathbb{E}[X_1]| \leq (b-a)$ p.s. donc $\mathbb{V}(X_1) \leq (b-a)^2$.

L'inégalité de Hoeffding permet d'améliorer cet intervalle car

$$\mathbb{P}_m\left[|\overline{X} - m| \geq t\right] \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

En posant

$$t = (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}$$

on obtient un second intervalle (non-exact également)

$$I_2 = \left]\overline{X}_n - (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\ ;\ \overline{X}_n + (b-a)\sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}\right[.$$

Les contributions de la taille de l'échantillon et de la longueur du support sont les mêmes, mais la contribution de α est nettement meilleure dans le second cas, lorsque α est proche de 0. En effet, quand α tend vers 0, le risque diminue et donc l'intervalle de confiance s'élargit. Mais il s'élargit beaucoup moins vite avec la seconde inégalité (à vitesse logarithmique) qu'avec la première (à la vitesse de $\sqrt{\alpha}$).

4.3.4 Composition par une fonction monotone

Si on a un IC au niveau $1-\alpha$ de θ noté $[\hat{a}, \hat{b}]$ et h est monotone, un IC au niveau $1-\alpha$ de $h(\theta)$ est $[h(\hat{a}), h(\hat{b})]$ si h est croissante, $[h(\hat{b}), h(\hat{a})]$ si h est décroissante.

Exemple :

Soit $(X_1, ..., X_n)$ un n-échantillon de loi $\mathcal{N}(m, \sigma^2)$.

Supposons m et σ^2 inconnus et intéressons-nous à l'estimation de $\sigma = \sqrt{\sigma^2}$.
On avait l'intervalle de confiance suivant pour σ^2 :

$$IC^{(3)}_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)S_n^2}{q^{\chi^2(n-1)}_{1-\frac{\alpha}{2}}}, \frac{(n-1)S_n^2}{q^{\chi^2(n-1)}_{\frac{\alpha}{2}}}\right].$$

On pose alors

$$IC_{1-\alpha}(\sigma) = \left[\sqrt{\frac{(n-1)S_n^2}{q^{\chi^2(n-1)}_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S_n^2}{q^{\chi^2(n-1)}_{\frac{\alpha}{2}}}}\right].$$

4.3.5 Inversion d'un test statistique

Les tests sont le sujet du chapitre suivant. Nous y verrons le lien entre test et région de confiance et une méthode pour déterminer des intervalles de confiance à partir des régions de rejet des tests.

4.4 Région de confiance asymptotique

4.4.1 Définition

Définition 23. *On se donne un modèle d'échantillonnage* $\left(\mathbb{X}^n, \mathfrak{F}^{\otimes n}, \left\{P_\theta^{\otimes n}, \theta \in \Theta\right\}\right)$. *Soit* $\alpha \in]0,1[$. *Une région de confiance asymptotique au niveau de confiance* $1-\alpha$ *pour l'estimation de* $\theta \in \Theta$ *notée* $RC_{1-\alpha}(\cdot, \theta)$ *ou simplement* $RC_{1-\alpha}(\theta)$ *(sans la dépendance en* X*, celle en* n *n'est jamais écrite) est une suite de fonctions* : $\mathbb{X}^n \to \mathcal{P}(\Theta')$ *avec* $\Theta \subseteq \Theta'$*, qui ne dépend pas de* θ*, telle que* $\{\theta \in RC_{1-\alpha}(X, \theta)\}$ *est mesurable pour tout* $\theta \in \Theta$*, tout* $n \geq 1$ *et tout* $X \sim P_\theta^{\otimes n}$ *et telle que :*

$$\forall \theta \in \Theta, \liminf_{n\to\infty} \mathbb{P}_\theta\left(\theta \in RC_{1-\alpha}(\theta)\right) \geq 1-\alpha. \tag{4.5}$$

Elle est exacte si il y a égalité et une vraie limite plutôt qu'une limite inférieure dans (4.5)*. Elle est plutôt appelée intervalle de confiance et notée* $IC_{1-\alpha}(X, \theta)$ *si* $\theta \subseteq \mathbb{R}$ *et* $RC_{1-\alpha}(x, \theta)$ *est un intervalle pour tout* $n \geq 1$ *et* $x \in \mathbb{X}^n$.

4.4.2 Méthode du pivot asymptotique

Définition 24. *Un pivot asymptotique $Z = Z(X_1, \dots, X_n, \theta)$ est suite de fonctions de l'échantillon $X = (X_1, \dots, X_n)$ et du paramètre θ telle que Z_n converge en loi vers une v.a. dont la loi est libre en θ, c'est à dire que sa loi ne dépend pas de θ.*

Soit $\widehat{\theta}_n$ un estimateur asymptotiquement normal de θ (ou plus généralement d'une fonction $h(\theta)$) :

$$\sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\sigma \sim \mathcal{N}(0, \sigma^2).$$

Alors un pivot asymptotique est donné par

$$Z_n = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\sigma} \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1)$$

Si σ est connu, on calcule le quantile $q^*_{1-\frac{\alpha}{2}}$ d'ordre $1-\alpha/2$ de la loi normale centrée réduite et l'on a alors, par définition de la convergence en loi,

$$\lim_{n\to\infty} \mathbb{P}_\theta\left(-q^*_{1-\frac{\alpha}{2}} \le Z_n \le q^*_{1-\frac{\alpha}{2}}\right) = \lim_{n\to\infty} \mathbb{P}_\theta\left(|Z_n| \le q^*_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

Un intervalle de confiance asymptotique de niveau $1-\alpha$ pour θ est alors donné par

$$IC_{1-\alpha}(\theta) = \left\{\tilde{\theta} : \left|\sqrt{n}\frac{\widehat{\theta}_n - \tilde{\theta}}{\sigma}\right| \le q^*_{1-\frac{\alpha}{2}}\right\} = \left[\widehat{\theta}_n - \frac{\sigma}{\sqrt{n}}q^*_{1-\frac{\alpha}{2}}, \widehat{\theta}_n + \frac{\sigma}{\sqrt{n}}q^*_{1-\frac{\alpha}{2}}\right].$$

4.4.3 Méthode de Wald

Lorsque le σ précédent n'est pas connu (par exemple car il dépend de θ), la méthode de Wald consiste à substituer dans la méthode précédente σ par un estimateur consistant $\widehat{\sigma}_n$ de l'écart type (par exemple par substitution si on possède un estimateur $\widehat{\theta}_n$ consistant et si $\theta \mapsto \sigma(\theta)$ est continue, en posant $\widehat{\sigma}_n = \sigma(\widehat{\theta}_n)$) et d'utiliser ensuite le LAC et le corollaire multiplicatif du lemme de Slutsky. Alors on aura :

$$\tilde{Z}_n = \frac{\sqrt{n}\left(\widehat{\theta}_n - \theta\right)}{\widehat{\sigma}_n} = Z_n \times \frac{\sigma}{\widehat{\sigma}_n} \xrightarrow[n\to\infty]{\mathcal{L}} Z \times 1 = Z \sim \mathcal{N}(0,1)$$

$\tilde{Z}_n$ est donc un pivot asymptotique utilisable pour construire un intervalle de confiance asymptotique.

Si $q^*_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{N}(0,1)$, alors l'intervalle de confiance asymptotique est

$$IC_{1-\alpha}(\theta) = \left[\widehat{\theta}_n - \frac{\widehat{\sigma}_n}{\sqrt{n}}q^*_{1-\frac{\alpha}{2}}, \widehat{\theta}_n + \frac{\widehat{\sigma}_n}{\sqrt{n}}q^*_{1-\frac{\alpha}{2}}\right].$$

Exemple :

On considère un n-échantillon de Bernoulli de paramètre θ à estimer. Le paramètre d'intérêt est la moyenne de chaque v.a. et un estimateur est donné par la moyenne empirique. Le théorème de la limite centrale montre que

$$\sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1)$$

D'après la consistance de $\overline{X}_n$ et la continuité de $x \mapsto \frac{\sqrt{\theta(1-\theta)}}{\sqrt{x(1-x)}}$, le LAC donne que

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}} \xrightarrow[n\to\infty]{\mathbb{P}} 1$$

qui est une constante, donc le corollaire multiplicatif du lemme de Slutsky s'applique et

$$\sqrt{n}\frac{\overline{X}_n-\theta}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}}=\sqrt{n}\frac{\overline{X}_n-\theta}{\sqrt{\theta(1-\theta)}}\times\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}}\xrightarrow[n\to\infty]{\mathcal{L}}Z\times 1=Z\sim\mathcal{N}(0,1)$$

Ainsi, $\sqrt{n}\frac{\overline{X}_n-\theta}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}}$ est un pivot asymptotique qui permet de construire l'intervalle de confiance asymptotique suivant :

$$IC^{(1)}_{1-\alpha}(\theta)=\left[\overline{X}_n-\frac{q^*_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\overline{X}_n(1-\overline{X}_n)}\ ;\ \overline{X}_n+\frac{q^*_{1-\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\overline{X}_n(1-\overline{X}_n)}\right]$$

où $q^*_{1-\frac{\alpha}{2}}$ est la quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

4.4.4 Méthode de stabilisation de la variance

On reprend un des exemples précédents :
Exemple :

Soit $(X_1,...,X_n)$ un n-échantillon d'une v.a. X de loi de Bernoulli de paramètre $\theta\in]0,1[$. Soit $\overline{X}_n$ la moyenne empirique de l'échantillon. D'après le théorème de la limite centrale,

$$\sqrt{n}\left(\overline{X}_n-\theta\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,\theta(1-\theta))$$

Posons $\psi(\theta)=2\arcsin\sqrt{\theta}$. Comme

$$\psi'(\theta)=\frac{1}{\sqrt{\theta(1-\theta)}},$$

Alors $J_\theta\Sigma_\theta J'_\theta=\psi'(\theta)^2\times\theta(1-\theta)=1$ et l'on a donc

$$\sqrt{n}\left(2\arcsin\sqrt{\overline{X}_n}-2\arcsin\sqrt{\theta}\right)\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,1).$$

Donc $\mathbb{P}_\theta\left(\sqrt{n}\left|2\arcsin\sqrt{\overline{X}_n}-2\arcsin\sqrt{\theta}\right|\le q^*_{1-\frac{\alpha}{2}}\right)\xrightarrow[n\to\infty]{}1-\alpha$, soit

$$\lim_{n\to\infty}\mathbb{P}_\theta\left(2\arcsin\sqrt{\overline{X}_n}-\frac{1}{\sqrt{n}}q^*_{1-\frac{\alpha}{2}}\le 2\arcsin\sqrt{\theta}\le 2\arcsin\sqrt{\overline{X}_n}+\frac{1}{\sqrt{n}}q^*_{1-\frac{\alpha}{2}}\right)=1-\alpha.$$

Un intervalle de confiance asymptotique est alors

$$IC^{(2)}_{1-\alpha}(\theta)=\left[\sin\left(\left(\arcsin\sqrt{\overline{X}_n}-\frac{1}{2\sqrt{n}}q^*_{1-\frac{\alpha}{2}})\right)\vee 0\right)^2,\sin\left(\left(\arcsin\sqrt{\overline{X}_n}+\frac{1}{2\sqrt{n}}q^*_{1-\frac{\alpha}{2}}\right)\wedge\frac{\pi}{2}\right)^2\right]$$

4.4.5 Inversion d'un test statistique

Comme dans la section 4.3.5.

4.5 Choix des quantiles

Une loi de probabilité absolument continue de fonction de répartition F est unimodale s'il existe x_M tel que F soit convexe avant x_M et concave après. x_M est alors le mode de F.

Proposition 10. *Soit f la densité du pivot Z. Si f est unimodale de mode x_M et vérifie, pour $a\le b$,*

- $\mathbb{P}(a\le Z\le b)=\int_a^b f(x)dx=1-\alpha$
- $f(a)=f(b)>0$
- $a<x_M<b$

Alors $[a,b]$ est l'intervalle de longueur minimale vérifiant la première condition

Ainsi, par exemple, $q_{1-\alpha/2}$ est le meilleur choix de quantile pour une loi symétrique. $q_{\alpha/2}$ et $q_{1-\alpha/2}$ ne sont pas optimaux pour une loi du chi deux par exemple, mais sont souvent utilisés par défaut.

Chapitre 5

Tests statistiques

5.1 Énoncé du problème, notion d'hypothèses

On réalise un test statistique lorsque l'on cherche à répondre à une question fermée sur la loi d'une variable que l'on observe.

Exemple :

On possède une balance de précision 0.01 kg. On sait que l'on pesait 62 kg trois semaines plus tôt, on cherche à savoir si l'on a pris du poids, et on pour cela on se pèse 10 fois. Comme souvent lorsque la variable à modéliser est continue, on va choisir un modèle gaussien. Plus précisément, le modèle statistique est le modèle d'échantillonnage $\left(\mathbb{R}^{10}, \mathcal{B}\left(\mathbb{R}^{10}\right), \left\{\mathcal{N}(\theta,(0.01)^2)^{\otimes 10}, \theta \in [62,\infty[\right\}\right)$, paramétré par $\theta \in \Theta = [62,\infty[$ qui représente donc le véritable poids. La traduction de la question posée en termes du modèle est : est-ce que $\theta > 62$ ou pas ? Autrement dit est-ce que $\theta \in \Theta_1 =]62,\infty[$ ou $\theta \in \Theta_0 = \{62\}$? On dit encore que l'on teste l'hypothèse nulle $H_0 : \theta \leq 62$ contre l'hypothèse alternative $H_1 : \theta > 62$.

C'est la même chose dans le cas général : on se place dans un modèle paramétré $(\mathbb{X}, \mathfrak{F}, \{P_\theta, \theta \in \Theta\})$. On dispose d'une partition de Θ en Θ_0 et Θ_1, et des deux sous-familles de lois associées $H_0 = \{P_\theta, \theta \in \Theta_0\}$, appelée hypothèse nulle, et $H_1 = \{P_\theta, \theta \in \Theta_1\}$, appelée hypothèse alternative.

Définition 25. *Une hypothèse est dite simple si elle contient une seule loi de probabilité, elle est dite composite sinon.*

On veut décider si la loi de la variable qu'on observe appartient à H_0 ou à H_1. Pour cela on contruit ce que l'on appelle un test statistique.

Définition 26. *Un test (pur) est une statistique $\phi : \mathbb{X} \longrightarrow \{0,1\}$.*

On dit que l'on rejette H_0 si $\phi(X) = 1$ (donc on pense que l'on est sous H_1), et que l'on conserve H_0 si $\phi(X) = 0$ (on pense que H_0 est vraie).

Dans la suite, on construira et manipulera des tests de la forme $\phi : x \mapsto \mathbb{1}_{T(x)\in R}$ où $T : \mathbb{X} \to \mathbb{R}$ est une statistique qu'on appelle statistique de test. Le nom "région de rejet" désigne selon le contexte et les ouvrages soit R, soit l'événement $\{T(X) \in R\}$ (il y a donc ambiguïté sur ce point de vocabulaire). Dans ce cours, on parlera de région de rejet pour R et d'événement de rejet pour $\{T(X) \in R\}$.

Construire un test revient donc à se donner une statistique de test et une région de rejet. Sauf qu'évidemment on ne va pas le faire n'importe comment, on veut des garanties statistiques sur ce que l'on fait. De la même façon que l'on veut un niveau de confiance sur les intervalles de confiance, on souhaite avoir confiance en nos décisions de rejet ou de conservation.

5.2 Erreur, risque, niveau, puissance

Définition 27.

— *On commet une erreur de première espèce si on rejette $\mathcal{H}_0$ à tort (on a rejeté alors que $\theta \in \Theta_0$). On parle aussi de faux positif ou de fausse découverte.*

— *On commet une erreur de deuxième espèce si on conserve $\mathcal{H}_0$ à tort (on n'a pas rejeté alors que $\theta \in \Theta_1$). On parle aussi de faux négatif.*

Définition 28.

— *Le risque de première espèce est la fonction qui à un $\theta \in \Theta_0$ associe la probabilité de commettre une erreur de première espèce avec le test $\phi(X)$ quand X est tirée selon P_θ, donc la probabilité de rejeter H_0 dans ce contexte :*

$$\alpha : \left\{ \begin{array}{rcl} \Theta_0 & \longrightarrow & [0,1] \\ \theta & \longmapsto & \mathbb{P}_\theta\left(\phi(X)=1\right) = \mathbb{P}_\theta\left(T(X) \in R\right) = \mathbb{E}_\theta\left[\phi(X)\right] \end{array} \right.$$

— *Le risque de seconde espèce est la fonction qui à un $\theta \in \Theta_1$ associe la probabilité de commettre une erreur de deuxième espèce avec le test $\phi(X)$ quand X est tirée selon P_θ, donc la probabilité de conserver H_0 dans ce contexte :*

$$\beta : \left\{ \begin{array}{rcl} \Theta_1 & \longrightarrow & [0,1] \\ \theta & \longmapsto & \mathbb{P}_\theta\left(\phi(X)=0\right) = \mathbb{P}_\theta\left(T(X) \notin R\right) = \mathbb{P}_\theta\left(T(X) \in R^{\mathsf{c}}\right) = 1 - \mathbb{E}_\theta\left[\phi(X)\right] \end{array} \right.$$

À partir du risque de seconde espèce β, on calcule la puissance γ du test, définie par $\gamma = 1-\beta$, de sorte que γ prolonge à Θ_1 la fonction α, puisque c'est la probabilité de (correctement) rejeter H_0, sous H_1.

Et l'on peut résumer les liens entre $\alpha(\cdot)$, $\beta(\cdot)$ et $\gamma(\cdot)$ dans le tableau ci-dessous :

Décision \Réalité	$\mathcal{H}_0$	$\mathcal{H}_1$
$\mathcal{H}_0$	$1-\alpha$	β
$\mathcal{H}_1$	α	$\gamma = 1-\beta$

Le problème est le suivant : on ne peut pas contrôler les deux risque en même temps : diminuer l'un augmente l'autre. En effet, réduire le risque de première espèce conduit à rejeter moins souvent et donc à augmenter le risque de seconde espèce, et inversement. Il faut donc en privilégier un, et ce sera le risque de première espèce. Étant donné qu'on va le contrôler, c'est-à-dire s'assurer qu'il est petit, on aura confiance en notre décision si on rejette H_0. Par contre le risque ne sera pas contrôlé (et même parfois inconnu) si on conserve H_0, ce ne sera pas une décision prise avec grande confiance, cela explique pour quoi on utilise le terme faible "conserver" plutôt que, par exemple, "accepter". Cela revient encore à considérer que H_0 est vraie *a priori* et que l'on va essayer de se convaincre (ou pas) à l'aide des observations, que ce n'est pas le cas. H_0 est l'hypothèse privilégiée, l'hypothèse par défaut, en ce sens qu'elle est supposée vérifiée tant que les observations ne conduisent pas à la rejeter.

Définition 29. *La taille α^* d'un test est*

$$\alpha^* = \sup_{\theta \in \Theta_0} \alpha(\theta) \tag{5.1}$$

Ce supremum existe (il est défini sur un ensemble non vide et borné) mais n'est pas toujours atteint. On a donc :

$$\forall \theta \in \Theta_0, \mathbb{P}_\theta\left(R\right) \leq \alpha^*.$$

Définition 30. *Le test est dit de niveau $\alpha \in]0,1[$ si $\alpha^* \leq \alpha$. Le niveau est exact si de plus $\alpha^* = \alpha$.*

La garantie statistique que l'on cherche à avoir est donc que le test soit d'un niveau α donné. Attention à l'ambiguïté : le symbole α désigne à la fois le niveau (fixé par l'utilisateur) recherché, et la fonction $\alpha(\cdot)$ risque de première espèce.

De la même façon que l'on préfère les intervalles de confiance exacts si on a le choix, on cherche à construire des tests de niveau exact si possible (et, comme pour les IC, c'est le plus souvent possible quand les lois manipulées sont continues, tandis que les lois discrètes le permettent rarement). La raison est la suivante : sous la contrainte que le test est de niveau α, donc de contrôle du risque de première espèce (on parle encore de condition de niveau), on cherche quand même à réduire le risque de seconde espèce, donc à augmenter les rejets, donc à augmenter la

taille de la région de rejet, donc à augmenter le risque de première espèce, qui ne peut augmenter que jusqu'à α.

Exemple : Retour à l'exemple de la pesée. On rappelle qu'on teste $H_0 : \theta = 62$ (hypothèse simple) contre $H_1 : \theta > 62$ (hypothèse composite). Et on cherche T et R tels que $\alpha(62) = \mathbb{P}_{62}(T_n(X) \in R) = \alpha$ (condition de niveau exact), avec $X = (X_1, \dots, X_n)$ avec $n = 10$. On propose comme stat de test

$$T_n(X) = \sqrt{n}\frac{\overline{X}_n - 62}{0.01}$$

qui est bien une statistique car elle ne dépend que de X et pas de θ. De plus, si $\theta = 62$, donc si on est dans H_0, sa loi est connue et c'est la loi $\mathcal{N}(0,1)$. Tandis que si $m > 62$ (on est sous H_1), $T_n(X) \sim \mathcal{N}\left(\sqrt{n}\frac{m-62}{0.01}, 1\right)$. Cette loi dépend de m qui est inconnu, mais on sait au moins qu'elle a tendance à prendre de plus grandes valeurs qu'une $\mathcal{N}(0,1)$, on dit encore qu'elle se décale à droite. Donc sous H_1, $T_n(X)$ a tendance à être grande, ce qui conduit à rejeter H_0 si on observe une grande $T_n(X)$, donc à poser R de la forme $]c_\alpha, \infty[$, avec c_α appelée valeur ou seuil critique, à déterminer (on dit aussi calibrer) en résolvant la condition de niveau $\mathbb{P}_{62}(T_n(X) \in R) = \alpha$ (d'où la dépendance en α). Et l'événement de rejet sera donc $\{T_n(X) > c_\alpha\}$. Or $\mathbb{P}_{62}(T_n(X) \in R) = \mathbb{P}(Z \sim \mathcal{N}(0,1) > c_\alpha) = \alpha$ implique que $c_\alpha = q_{1-\alpha}^{\mathcal{N}(0,1)}$ ce qui achève la construction du test.

Enfin, deux tests de même niveau se comparent sur leur puissance ou, de façon équivalente, leur erreur de seconde espèce.

5.3 Constructions fréquentes, hypothèse nulle simple

On considère les quatre cas fréquents dans le cas où H_0 est simple : $\Theta_0 = \{\theta_0\}$, et où on dispose d'une statistique T dont on connaît la loi sous H_0, notée $\mathcal{L}_0$. On ne considère pas ici que $\mathcal{L}_0$ est continue, on reste dans le cas général. À noter qu'on ne peut rien faire si on ne connaît pas la loi de $T(X)$ sous H_0, ce pré-requis fait partie pour beaucoup de la définition d'une statistique de test.

1. si, sous H_1, la loi de T se décale à droite : elle a tendance à prendre de plus grandes valeurs que sous H_0. Alors on prendra $R =]c_\alpha, \infty[$ et la calibration par la condition de niveau donne $c_\alpha = q_{1-\alpha}^{\mathcal{L}_0}$.
2. si, sous H_1, la loi de T se décale à gauche : elle a tendance à prendre de plus petites valeurs que sous H_0. Alors on prendra $R =]-\infty, c_\alpha[$ et la calibration par la condition de niveau donne $c_\alpha = \max\{c : \mathbb{P}_{Z\sim\mathcal{L}_0}(Z < c_\alpha) \leq \alpha\}$.
3. si, sous H_1, la loi de T se décale de deux côtés : elle a tendance à prendre de plus grandes valeurs ou de plus petites que sous H_0. Alors on prendra $R =]-\infty, c_{1,\alpha}[\cup]c_{2,\alpha}, \infty[$ et la calibration par la condition de niveau donne $c_{2,\alpha} = q_{1-\frac{\alpha}{2}}^{\mathcal{L}_0}$ et $c_{1,\alpha} = \max\{c : \mathbb{P}_{Z\sim\mathcal{L}_0}(Z < c_\alpha) \leq \frac{\alpha}{2}\}$.
4. même cas que le précédent, mais on suppose en plus que $\mathcal{L}_0$ est symétrique (par rapport à 0). Alors on prendra $R =]-\infty, -c_\alpha[\cup]c_\alpha, \infty[$ (d'où l'événement de rejet $\{|T(X)| > c_\alpha\}$) et la calibration par la condition de niveau donne $c_\alpha = q_{1-\frac{\alpha}{2}}^{\mathcal{L}_0}$.

En effet, pour le premier cas par exemple, on souhaite, par la condition de niveau, avoir $\mathbb{P}_{\theta_0}(T(X) > c_\alpha) = \mathbb{P}_{Z\sim\mathcal{L}_0}(Z > c_\alpha) \leq \alpha$ soit encore $F_{\mathcal{L}_0}(c_\alpha) = \mathbb{P}_{Z\sim\mathcal{L}_0}(Z \leq c_\alpha) \geq 1-\alpha$, et comme on veut minimiser le risque de seconde espèce sous cette contrainte on cherche à maximiser R et donc minimiser c_α, ce qui conduit à poser $c_\alpha = \min\{c : \mathbb{P}_{Z\sim\mathcal{L}_0}(Z \leq c) \geq 1-\alpha\} = q_{1-\alpha}^{\mathcal{L}_0}$ par définition de la fonction quantile.

À noter : le second cas se ramène au premier en prenant $-T$ à la place de T. On en déduit l'égalité $\max\{c : \mathbb{P}_{Z\sim\mathcal{L}_0}(Z < c_\alpha) \leq \alpha\} = -q_{1-\alpha}^{-\mathcal{L}_0}$, où $-\mathcal{L}_0$ désigne la loi de $-T(X)$ sous H_0. On peut montrer que c'est, bien plus simplement, $q_\alpha^{\mathcal{L}_0}$ si $\mathcal{L}_0$ est continue ou si elle est discrète et $F_{\mathcal{L}_0}(q_\alpha^{\mathcal{L}_0}) > \alpha$. Si $\mathcal{L}_0$ est discrète, $F_{\mathcal{L}_0}(q_\alpha^{\mathcal{L}_0}) = \alpha$, et le support de $\mathcal{L}_0$ est discret (pas seulement dénombrable), alors on peut montrer que $\max\{c : \mathbb{P}_{Z\sim\mathcal{L}_0}(Z < c_\alpha) \leq \alpha\}$ est l'élément suivant $q_\alpha^{\mathcal{L}_0}$ dans le support de $\mathcal{L}_0$: $\min\{x \in \text{supp}(\mathcal{L}_0) : x > q_\alpha^{\mathcal{L}_0}\}$.

À noter enfin : le choix de distribuer l'erreur en $\alpha/2$ à gauche et $\alpha/2$ à droite dans le troisième cas n'a rien d'optimal (au sens de la réduction du risque de seconde espèce), là aussi on peut

faire l'analogie avec les intervalles de confiance où le choix est le plus souvent arbitraire (pas dans le quatrième cas).

Définition 31. *Dans les deux premiers cas, le test est dit unilatéral, dans le troisième, il est bilatéral. Le quatrième est généralement considéré comme bilatéral aussi mais il devient unilatéral si on change T pour $|T|$ comme stat de test.*

Exemple : On se donne un échantillon $X = (X_1, \dots, X_n)$ de loi commune $\mathcal{E}(\lambda)$ et on cherche à tester $H_0 : \lambda = 3$ contre $H_1 : \lambda > 3$. Comme souvent, on commence par estimer le paramètre sur lequel porte le test. Un EMM, et l'EMV, est $\hat{\lambda}_n = \frac{1}{\overline{X}_n}$, ce qui nous conduit à considérer $T_n(X) = \sum_{i=1}^n X_i$ qui suit une loi Gamma(n, λ) par indépendance. Donc sous H_0, $T_n(X) \sim$ Gamma$(n, 3)$. Sous H_1, la loi de $T_n(X)$ se décale à gauche (à cause de l'inverse, attention à cette subtilité), on est donc dans le deuxième cas ci-dessus, et on rejette H_0 si $T_n(X) < q_\alpha^{\text{Gamma}(n,3)}$.

Exemple : Même cas que précédemment mais où cette fois on teste $H_0 : \lambda = 2$ contre $H_1 : \lambda \neq 2$. Cette fois on effectue un test bilatéral avec une loi sous H_0 qui est la loi Gamma$(n, 2)$ (non-symétrique) et donc on pose

$$\phi(X) = \mathbb{1}_{T_n(X) < q_{\frac{\alpha}{2}}^{\text{Gamma}(n,2)} \text{ ou } T_n(X) > q_{1-\frac{\alpha}{2}}^{\text{Gamma}(n,2)}}.$$

Exemple : On reprend l'exemple de la pesée mais où cette fois on ne connaît pas la précision de la balance. On pose alors comme statistique de test

$$T_n(X) = \sqrt{n}\frac{\overline{X}_n - 62}{\sqrt{S_n^2}}$$

avec S_n^2 l'estimateur non-biaisé de la variance. Sous H_0, on sait que $T_n(X)$ suit une loi $\mathcal{T}(n-1)$ d'où l'événement de rejet $\{T_n(X) > q_{1-\alpha}^{\mathcal{T}(n-1)}\}$.

Exemple : Toujours le même exemple où cette fois on s'intéresse justement à la précision de la balance. On teste $H_0 : \sigma^2 = (0.01)^2$ contre $H_1 : \sigma^2 > (0.01)^2$. On sait que $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$. On pose comme stat de test $T_n(X) = \frac{(n-1)S_n^2}{(0.01)^2}$. Sous H_0, T_n suit donc la loi $\chi^2(n-1)$. De plus on a $T_n(X) = \frac{(n-1)S_n^2}{\sigma^2} \times \frac{\sigma^2}{(0.01)^2}$ donc sous H_1 T_n est le produit d'une $\chi^2(n-1)$ et d'un nombre > 1 : elle se décale à droite et le test est $\phi(X) = \mathbb{1}_{\frac{(n-1)S_n^2}{(0.01)^2} > q_{1-\alpha}^{\chi^2(n-1)}}$.

5.4 Constructions fréquentes, hypothèse nulle composite

Le théorème de Lehmann affirme, de façon informelle, que si le rapport de vraisemblance, dans un modèle où $\Theta \subseteq \mathbb{R}$, peut s'écrire comme une fonction monotone d'une certaine statistique T, alors le test avec hypothèse nulle composite peut se construire comme si l'hypothèse nulle était simple, en utilisant T comme statistique de test.

Théorème 12 (Théorème de Lehmann (informel))**.** *On suppose $\Theta \subseteq \mathbb{R}$. On suppose qu'il existe une statistique T telle que soit :*

1. *pour tout $\theta \leq \theta'$, il existe une fonction croissante $g_{\theta,\theta'}$ telle que pour tout $x \in \mathbb{X}$, $\frac{L(\theta';x)}{L(\theta;x)} = g_{\theta,\theta'}(T(x))$ (cas croissant),*

soit :

2. *pour tout $\theta \leq \theta'$, il existe une fonction décroissante $g_{\theta,\theta'}$ telle que pour tout $x \in \mathbb{X}$, $\frac{L(\theta';x)}{L(\theta;x)} = g_{\theta,\theta'}(T(x))$ (cas décroissant).*

Alors, pour tout $\theta_0 \in \Theta$,

1. *le test de $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$ se construit avec T comme statistique de test comme le test $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$. De plus la région de rejet est de la forme $]c, \infty[$ dans le cas croissant et $]-\infty, c[$ dans le cas décroissant.*

2. *le test de* $H_0 : \theta \geq \theta_0$ *vs* $H_1 : \theta < \theta_0$ *se construit avec* T *comme statistique de test comme le test* $H_0 : \theta = \theta_0$ *vs* $H_1 : \theta < \theta_0$. *De plus la région de rejet est de la forme* $]-\infty, c[$ *dans le cas croissant et* $]c, \infty[$ *dans le cas décroissant.*

On remarque encore une fois que le cas décroissant se ramène au cas croissant si on pose $T' = -T$. C'est en dehors du cadre de ce cours, mais le théorème de Lehmann dans son intégralité comprend aussi un résultat d'optimalité du test avec T, dans un certain sens.

Exemple : On reprend l'exemple exponentiel mais où cette fois on teste $H_0' : \lambda \leq 3$ contre $H_1' : \lambda > 3$ au lieu de $H_0 : \lambda = 3$ contre $H_1 : \lambda > 3$. Le théorème de Lehmann dit que le test ne change pas et on rejette H_0 si $T_n(X) < q_\alpha^{\text{Gamma}(n,3)}$. En effet soit $\lambda \leq \lambda'$,

$$\begin{aligned}\frac{L_n(\lambda'; X)}{L_n(\lambda; X)} &= \frac{\prod_{i=1}^n (\lambda' \exp(-\lambda' X_i))}{\prod_{i=1}^n (\lambda' \exp(-\lambda X_i))} \\ &= \left(\frac{\lambda'}{\lambda}\right)^n \exp\left((\lambda - \lambda') \sum_{i=1}^n X_i\right) \\ &= \left(\frac{\lambda'}{\lambda}\right)^n \exp\left((\lambda - \lambda') T_n(X)\right)\end{aligned}$$

est bien une fonction décroissante de $T_n(X)$ vu que $\lambda - \lambda' \leq 0$.

Exemple : Dans le modèle d'échantillonnage de Bernoulli où $X_1, \ldots, X_n \sim \mathcal{B}(p)$, on teste $H_0 : p \geq \frac{1}{2}$ contre $H_1 : p < \frac{1}{2}$. Pour $p \geq p'$,

$$\begin{aligned}\frac{L_n(p'; X)}{L_n(p; X)} &= \frac{(p')^{\sum_{i=1}^n X_i}(1-p')^{n-\sum_{i=1}^n X_i}}{p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}} \\ &= \left(\frac{1-p'}{1-p}\right)^n \left(\frac{p'}{1-p'}\frac{1-p}{p}\right)^{\sum_{i=1}^n X_i}.\end{aligned}$$

Or $p \mapsto \frac{p}{1-p} = \frac{1}{\frac{1}{p}-1}$ est une fonction croissante de p donc $\frac{p'}{1-p'}\frac{1-p}{p} \geq 1$ et enfin $\frac{L_n(p';X)}{L_n(p;X)}$ est une fonction croissante de $S_n(X) = \sum_{i=1}^n X_i$. On va donc utiliser $S_n(X)$ comme stat de test et construire le test comme si on testait $H_0 : p = \frac{1}{2}$ contre $H_1 : p < \frac{1}{2}$, donc on rejette si $S_n(X) < c_\alpha$ avec $c_\alpha = \max\left\{c : F^-_{\mathcal{B}(n,\frac{1}{2})}(c) \leq \alpha\right\}$. Donc $c_\alpha = q_\alpha^{\mathcal{B}(n,\frac{1}{2})}$ si α n'est pas atteint par $F_{\mathcal{B}(n,\frac{1}{2})}$, et $c_\alpha = q_\alpha^{\mathcal{B}(n,\frac{1}{2})} + 1$ si il l'est.

À noter, dans cet exemple discret, le test n'est jamais de taille exactement α (donc le niveau n'est pas exact), sauf dans le rare cas où α est atteint par $F_{\mathcal{B}(n,\frac{1}{2})}$ (il y a seulement $n-1$ tels α). En effet,

$$\begin{aligned}\alpha^* &= \sup_{p \geq \frac{1}{2}} \alpha(p) \\ &= \sup_{p \geq \frac{1}{2}} \mathbb{P}_p\left(S_n(X) < c_\alpha\right) \\ &= \mathbb{P}_{\frac{1}{2}}\left(S_n(X) < c_\alpha\right) \\ &= F^-_{\mathcal{B}(n,\frac{1}{2})}(c_\alpha)\end{aligned}$$

avec, si α n'est pas atteint par $F_{\mathcal{B}(n,\frac{1}{2})}$, $c_\alpha = q_\alpha^{\mathcal{B}(n,\frac{1}{2})}$ et donc $\alpha^* = F^-_{\mathcal{B}(n,\frac{1}{2})}\left(q_\alpha^{\mathcal{B}(n,\frac{1}{2})}\right) = F_{\mathcal{B}(n,\frac{1}{2})}\left(q_\alpha^{\mathcal{B}(n,\frac{1}{2})} - 1\right) < \alpha$ (on utilise le fait que $\mathcal{B}\left(n, \frac{1}{2}\right)$ est à valeurs entières). Et, si α est atteint par $F_{\mathcal{B}(n,\frac{1}{2})}$, $c_\alpha = q_\alpha^{\mathcal{B}(n,\frac{1}{2})} + 1$ et donc $\alpha^* = F^-_{\mathcal{B}(n,\frac{1}{2})}\left(q_\alpha^{\mathcal{B}(n,\frac{1}{2})} + 1\right) = F_{\mathcal{B}(n,\frac{1}{2})}\left(q_\alpha^{\mathcal{B}(n,\frac{1}{2})}\right) = \alpha$.

5.5 p-valeur

On définit une p-valeur (p-value en anglais) comme "probabilité de réaliser sous H_0 un événement au moins aussi extrême que celui observé". Elle dépend donc de l'observation : c'est une statistique $P : \mathbb{X} \to [0,1]$, et si on la compose par l'observation, c'est donc une variable aléatoire.

On propose trois constructions qui répondent à l'intuition précédente. On considère qu'on a une région de rejet R et une statistique de test T.

1. Si $R =]c, \infty[$ on peut définir une p-valeur comme :

$$\begin{aligned} P(x) &= \sup_{\theta_0 \in \Theta_0} \mathbb{P}_{Z \sim P_{\theta_0}} \left(T(Z) \geq T(x) \right) \\ &= \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T^{-1}([T(x), \infty[)). \end{aligned}$$

   Le conditionnement en X est souvent omis dans la première notation. Il signifie qu'on ne calcule la probabilité qu'en Z, comme si X était fixé. C'est bien une statistique car une fonction mesurable de X.

2. Si $R =]-\infty, c[$ on peut définir une p-valeur comme :

$$\begin{aligned} P(x) &= \sup_{\theta_0 \in \Theta_0} \mathbb{P}_{Z \sim P_{\theta_0}} \left(T(Z) \leq T(x) \right) \\ &= \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T^{-1}(]-\infty, T(x)])). \end{aligned}$$

3. Si $R =]-\infty, c_1[\cup]c_2, \infty[$, on peut définir une p-valeur comme :

$$\begin{aligned} P(x) &= \sup_{\theta_0 \in \Theta_0} 2 \min \left(\mathbb{P}_{Z \sim P_{\theta_0}} \left(T(Z) \leq T(x) \right), \mathbb{P}_{Z \sim P_{\theta_0}} \left(T(Z) \geq T(x) \right) \right) \\ &= \sup_{\theta_0 \in \Theta_0} 2 \min \left(P_{\theta_0}(T^{-1}(]-\infty, T(x)])), P_{\theta_0}(T^{-1}([T(x), \infty[)) \right). \end{aligned}$$

Remarque : on n'a pas vraiment besoin de telle ou telle forme de région de rejet pour que ces trois définitions aient un sens et soient valides (au sens de la proposition ci-dessous). C'est juste que chacune des 3 constructions satisfait l'intuition donnée au début si la forme de la région correspond.

Ci-dessous la propriété fondamentale qu'on demande à toute p-value.

Proposition 11. *Si $X \sim P_\theta$ et $\theta \in \Theta_0$ et $P(\cdot)$ est définie dans un des cas ci-dessus, alors $P(X)$ est super-uniforme : sa fonction de répartition est inférieure ou égale à celle d'une loi $\mathcal{U}([0,1])$:*

$$\forall x \in \mathbb{R},\ \mathbb{P}\left(p(X) \leq x\right) \leq \mathbb{P}\left(U \leq x\right) = 0 \vee (x \wedge 1) \tag{5.2}$$

Démonstration. Uniquement pour la première construction. $P(X) \in [0,1]$ presque sûrement. donc il suffit de vérifier (5.2) uniquement pour $x \in [0,1[$. Soit $x \in]0,1[$, on traitera le cas $x = 0$ en dernier.

$$\begin{aligned} \mathbb{P}\left(P(X) \leq x\right) &= \mathbb{P}\left(\sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T^{-1}([T(X), \infty[)) \leq x \right) \\ &= \mathbb{P}\left(\bigcap_{\theta_0 \in \Theta_0} \{ P_{\theta_0}(T^{-1}([T(X), \infty[)) \leq x \} \right) \\ &\leq \mathbb{P}\left(P_{\theta_0}(T^{-1}([T(X), \infty[)) \leq x \right), \end{aligned}$$

vu que $\theta \in \Theta_0$. Soit F la fonction de répartition de $T(X)$, elle est càdlàg, soit F^- la fonction limite à gauche associée : $F^-(x) = \lim_{\substack{\varepsilon \to 0 \\ \varepsilon > 0}} F(x-\varepsilon)$. Alors

$$\begin{aligned}\mathbb{P}\left(P_{\theta_0}(T^{-1}([T(X),\infty[)) \leq x\right) &= \mathbb{P}\left(1 - P_\theta(T^{-1}(]-\infty, T(X)[)) \leq x\right) \\ &= \mathbb{P}\left(1 - x \leq F^-(T(X))\right) \\ &= \mathbb{P}\left(T(X) \in (F^-)^{-1}([1-x,1])\right).\end{aligned}$$

F^- est croissante, de limite 1 en ∞ et 0 en $-\infty$, et $0 < 1-x < 1$, donc $(F^-)^{-1}([1-x,1])$ est un intervalle de la forme $]a,\infty[$ ou $[a,\infty[$, donc on distingue deux cas. Si a est dedans alors

$$\begin{aligned}\mathbb{P}\left(P_\theta(T^{-1}([T(X),\infty[)) \leq x\right) &= \mathbb{P}\left(T(X) \geq a\right) \\ &= 1 - F^-(a) \\ &\leq 1 - (1-x) \text{ car } a \in (F^-)^{-1}([1-x,1]) \\ &= x.\end{aligned}$$

Sinon,

$$\begin{aligned}\mathbb{P}\left(P_\theta(T^{-1}([T(X),\infty[)) \leq x\right) &= \mathbb{P}\left(T(X) > a\right) \\ &= 1 - F(a)\end{aligned}$$

Mais $F(a)$ est la limite à droite de F^- en a : $F(a) = \lim_{\substack{\varepsilon \to 0 \\ \varepsilon > 0}} F^-(a+\varepsilon)$ avec $a+\varepsilon \in (F^-)^{-1}([1-x,1])$ donc $F(a) \geq 1-x$ par passage à la limite et donc $\mathbb{P}\left(P_\theta(T^{-1}([T(X),\infty[)) \leq x\right) \leq x$ aussi.

Maintenant qu'on a (5.2) pour tout $x \in]0,1[$, on fait $x \to 0$ et on se sert de la continuité à droite de F pour conclure avec le cas $x = 0$. □

Exercice : traiter les deux autres constructions de p-valeur.

On peut démontrer (à titre d'exercice) que $P(X)$ suit exactement une loi uniforme sur $[0,1]$ si $\Theta_0 = \{\theta_0\}$ et la loi de $T(X)$ est continue pour $X \sim P_{\theta_0}$.

Ce résultat a une conséquence importante, c'est que la p-valeur peut elle aussi être utilisée comme statistique de test :

Corollaire 2. *Le test $\mathbb{1}_{P(\cdot) \leq \alpha}$ est de niveau α.*

Démonstration. Pour tout $\theta \in \Theta_0$,

$$\alpha(\theta) = \mathbb{P}_{X \sim P_\theta}\left(P(X) \leq \alpha\right) \leq \alpha \text{ par (5.2).}$$

□

Si il existe $\theta \in \Theta_0$ tel que $p(X) \sim \mathcal{U}([0,1])$ si $X \sim P_\theta$ alors le test est même de taille α.

On voit là un bénéfice de l'approche par p-valeur : si on change le niveau du test, pas besoin de recalculer un quantile.

Exemple :

Exemples gaussiens. Dans le modèle $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{N}(m,1), m \in \mathbb{R}\})$, où $H_0 = \{\mathcal{N}(m,1), m \leq 0\}$ et $H_1 = \{\mathcal{N}(m,1), m > 0\}$, en utilisant comme statistique de test X elle-même,

$$\begin{aligned}P(x) &= \sup_{m \leq 0} P_m([x,\infty[) \\ &= \sup_{m \leq 0} \mathbb{P}_{Z \sim \mathcal{N}(m,1)}\left(Z \geq x\right) \\ &= \sup_{m \leq 0} (1 - \Phi(x-m)) = 1 - \Phi(x)\end{aligned}$$

et rejeter si $p(X) \leq \alpha$ est même équivalent à rejeter si $T(X) > q_{1-\alpha}^{\mathcal{N}(0,1)}$ (Exercice). Dans le modèle d'échantillonnage $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathcal{N}(m,1)^{\otimes n}, m \in \mathbb{R}\})$, où $H_0 = \{\mathcal{N}(m,1)^{\otimes n}, m \leq 0\}$ et

$H_1 = \{\mathcal{N}(m,1)^{\otimes n}, m > 0\}$, en utilisant comme statistique de test $T_n(X) = \sqrt{n}\frac{\overline{X}_n - 0}{\sqrt{1}} = \sqrt{n}\overline{X}_n$, on peut montrer de même que $P_n(X) = 1 - \Phi(T_n(X))$.

Exemple :

Retour à l'exemple binomial. Alors $P_n(X) = F_{\mathcal{B}\left(10,\frac{1}{2}\right)}(S_n(X))$. Avec $S_n(X) = \sum_{i=1}^n X_i$.

Dans de nombreux contextes de tests d'hypothèse simple, on peut montrer que, si la construction de $\alpha \mapsto R_\alpha$ est la "bonne", au sens où la construction choisie de la région de rejet correspond à la forme listée avec la construction de la p-valeur dans les définitions plus haut, le test de base et le test construit avec la p-valeur sont équivalents.

Proposition 12. *On suppose que pour tout $\theta \in \Theta, X \sim P_\theta$, la loi $\mathcal{L}_{T,\theta}$ de $T(X)$ est continue, ou que pour tout $\theta \in \Theta, X \sim P_\theta$, la loi $\mathcal{L}_{T,\theta}$ de $T(X)$ est discrète à support discret (pas seulement dénombrable, ce qui implique que le support n'a pas de point d'accumulation). On suppose que H_0 est simple : $H_0 = \{P_{\theta_0}\}$. Alors,*

1. *Pour tout $\theta \in \Theta$, $\mathbb{1}_{T(x) > q_{1-\alpha}^{\mathcal{L}_{T,\theta_0}}} = \mathbb{1}_{P(x) \leq \alpha}$ pour $P(\cdot) = \mathbb{P}_{Z \sim P_{\theta_0}}(T(Z) \geq T(\cdot))$, P_θ-presque partout. Autrement dit pour toute $X \sim P_\theta$, $\mathbb{1}_{T(X) > q_{1-\alpha}^{\mathcal{L}_{T,\theta_0}}} = \mathbb{1}_{P(X) \leq \alpha}$ presque sûrement.*
2. *Pour tout $\theta \in \Theta$, $\mathbb{1}_{T(x) < -q_{1-\alpha}^{\mathcal{L}_{-T,\theta_0}}} = \mathbb{1}_{P(x) \leq \alpha}$ pour $P(\cdot) = \mathbb{P}_{Z \sim P_{\theta_0}}(T(Z) \leq T(\cdot))$, P_θ-presque partout. Autrement dit pour toute $X \sim P_\theta$, $\mathbb{1}_{T(X) < -q_{1-\alpha}^{\mathcal{L}_{-T,\theta_0}}} = \mathbb{1}_{P(X) \leq \alpha}$ presque sûrement.*
3. *Pour tout $\theta \in \Theta$, $\mathbb{1}_{T(x) < -q_{1-\frac{\alpha}{2}}^{\mathcal{L}_{-T,\theta_0}} \text{ ou } T(x) > q_{1-\frac{\alpha}{2}}^{\mathcal{L}_{T,\theta_0}}} = \mathbb{1}_{P(x) \leq \alpha}$ pour $P(\cdot) = 2\min\left(\mathbb{P}_{Z \sim P_{\theta_0}}(T(Z) \leq T(\cdot)), \mathbb{P}_{Z \sim$ P_θ-presque partout. Autrement dit pour toute $X \sim P_\theta$, $\theta \in \Theta$, $\mathbb{1}_{T(X) < -q_{1-\frac{\alpha}{2}}^{\mathcal{L}_{-T,\theta_0}} \text{ ou } T(X) > q_{1-\frac{\alpha}{2}}^{\mathcal{L}_{T,\theta_0}}} =$ $\mathbb{1}_{P(X) \leq \alpha}$ presque sûrement.*

Dans le contexte de la proposition, si on fait varier le niveau souhaité α et qu'on note donc la région de rejet associée R_α, la p-valeur s'interprète donc aussi comme le plus petit niveau α' tel que $T(X) \in R_{\alpha'}$. C'est la définition qui est adoptée par certains auteurs, mais elle pose problème vu que l'équivalence n'est pas vraie tout le temps (on peut construire des contre-exemples avec par exemple des lois hybrides continues-discrètes) et que dans la théorie des tests multiples (au-delà du cadre de ce cours) cela correspond à la p-valeur ajustée qui est différente de la p-valeur.

Exemple :

Le ministère de la santé étudie régulièrement la nécessité de prendre des mesures contre la consommation d'alcool et l'efficacité de ces mesures. L'Insee fournit à cet effet des données annuelles de consommation moyenne d'alcool par personne et par jour. En 1991, la loi Évin interdit la publicité sur les boissons alcoolisées et lance une campagne de sensibilisation sous forme de spots publicitaires. Avant la loi Évin, la consommation d'alcool moyenne chez les personnes de plus de 15 ans était de 35 g par jour. L'objectif premier de la loi était de baisser cette consommation journalière moyenne à 33 g. En se basant sur l'observation des consommations moyennes de 1991 à 1994, le ministère a fixé la règle de décision suivante : si la moyenne des consommations journalières sur ces quatre années est supérieures à 34.2 g, alors les mesures prises ont été inefficaces.

On suppose que les données recueillies sont des réalisations de v.a. supposées i.i.d. et de loi normale $\mathcal{N}(\theta, \sigma^2)$ avec $\sigma = 2$. On note X_i, $1 \leq i \leq n$ avec $n = 4$, les moyennes de consommation journalières pour chacune des quatre années de l'enquête et l'on note $\overline{X}$ la moyenne empirique des X_i.

Les hypothèses à tester sont les suivantes :

$$\begin{cases} H_0 : \ \theta = 33 \\ H_1 : \ \theta = 35 \end{cases} \tag{5.3}$$

L'hypothèse nulle est $\theta = 33$ car on part du principe que la politique ait été efficace. On suppose donc l'effet significatif *a priori* et l'erreur que l'on va pouvoir contrôler est l'erreur d'affirmer que la campagne a été inefficace alors qu'elle l'était (choix très discutable, qui découle sans doute d'une politique visant à économiser le coût d'une seconde campagne).

On choisit $\overline{X}_n$ comme statistique pour mesurer l'effet moyen de la consommation journalière. On connaît sa loi sous H_0, c'est $\mathcal{N}(33, 1)$, et sous H_1 sa loi se décale à droite.

On voit que le ministère a choisi d'utiliser le test $\mathbb{1}_{\overline{X}_n > 34.2}$.

Nous commençons par calculer la taille α du test, en notant $s = 34.2$ le seuil choisi par le ministère (arbitrairement ?) :

$$\alpha(33) = \mathbb{P}_{33}(\overline{X}_n > s) = \mathbb{P}_{33}\left(\overline{X}_n - 33 > s - 33\right) = 1 - \Phi(34.2 - 33) \approx 0.1151,$$

où Φ désigne toujours la fonction de répartition de la loi normale centrée réduite.

Le risque de seconde espèce est

$$\beta(35) = \mathbb{P}_{35}(\overline{X}_n \leq s) = \mathbb{P}_{35}(\overline{X}_n - 35 \leq s - 35) = \Phi(s - 35) \approx 0.212.$$

Les deux risques du seuil proposé par le ministère sont donc plutôt élevés.

La valeur du niveau α par défaut est souvent 5%. Le seuil correspondant, que l'on notera s_α, vérifie :

$$\begin{aligned}
\mathbb{P}_{33}\left(\overline{X}_n > s_\alpha\right) = 0.05 = \alpha &\iff 1 - \mathbb{P}_{33}\left(\overline{X}_n - 33 < s^* - 33\right) = 0.05 \\
&\iff \Phi(s_\alpha - 33) = 0.95 \\
&\iff s^* = 33 + q_{1-\alpha}^{\mathcal{N}(0,1)} \simeq 34.65.
\end{aligned}$$

Ainsi le test défini par $\phi(x) = \mathbb{1}_{\overline{x}_n > 34.65}$ est donc de niveau (exact) $\alpha = 5\%$.

Si l'on a observé $x = (34.7\ ,\ 34.4\ ,\ 33.7\ ,\ 33.3)$, alors la moyenne est $\overline{x}_n = 34.025$ et la p-valeur du test (pour cette observation) est

$$P(\overline{x}_n) = \mathbb{P}_{33}(\overline{X}_n \geq 34.025) = 1 - \Phi(1.025) \approx 0.153. \tag{5.4}$$

0.153 est le niveau minimum auquel on va rejeter $\mathcal{H}_0$ pour ces observations. 0.05 est plus petit que cette p-valeur et effectivement, on conserve $\mathcal{H}_0$ à ce niveau de 5% car la moyenne observée de 34.025 est bien inférieure à 34.645. Quand on conserve, le risque d'erreur est le risque de seconde espèce et il n'est pas contrôlé. On peut le calculer ici, comme avant mais avec s_α au lieu de s :

$$\beta(35) = \mathbb{P}_{35}(\overline{X}_n \leq s_\alpha) = \mathbb{P}_{35}(\overline{X}_n - 35 \leq s - 35) = \Phi(34.65 - 35) \approx 0.36.$$

Le risque de seconde espèce est plus élevé que pour le test du ministère. C'est normal, vu qu'on a réduit la zone de rejet pour s'assurer du contrôle du risque de première espèce au niveau 5% (on dit qu'on est plus conservatif).

5.6 Tests asymptotiques

5.6.1 Définition

Parfois on ne peut pas construire de test de niveau α car on n'a pas sous la main de statistique de test dont on connaît assez bien la loi pour construire une région de rejet associée qui vérifie la condition de niveau. Cependant, si on connaît la loi asymptotique de la statistique de test, on peut construire un test dit asymptotique, dont le niveau est asymptotique.

Définition 32. *Le test est de niveau asymptotique α si on est dans un modèle d'échantillonnage et $\limsup_{n\to\infty} \alpha^* \leq \alpha$. Le niveau est dit exact si en plus on a $\lim_{n\to\infty} \alpha^* = \alpha$.*

Exemple : On reprend le modèle d'échantillonnage de Bernoulli où $X_1, \dots, X_n \sim \mathcal{B}(p)$, on teste $H_0 : p \geq \frac{1}{2}$ contre $H_1 : p < \frac{1}{2}$. avec $S_n(X) = \sum_{i=1}^n X_i$ la stat de test et le test qui se construit comme le test de $H_0 : p = \frac{1}{2}$ contre $H_1 : p < \frac{1}{2}$ par Lehmann. On a vu que l'on rejetait si $S_n(X) < \max\left\{c : F^-_{\mathcal{B}\left(n,\frac{1}{2}\right)}(c) \leq \alpha\right\}$, mais cette valeur devient très compliquée à déterminer quand n devient grand. On peut alors se tourner vers la construction d'un test de niveau asymptotique α grâce à la normalité asymptotique de l'estimateur de p. Sous H_0, $p = \frac{1}{2}$ et on a, par le TLC appliqué à l'échantillon L^2 $(X_1, \dots, X_n)$,

$$\sqrt{n}\frac{\frac{1}{n}S_n(X) - \frac{1}{2}}{\sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)}} \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1)$$

soit

$$\sqrt{n}\left(\frac{2}{n}S_n(X) - 1\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1).$$

On a toujours que la loi de $S_n(X)$ se décale à gauche sous H_1, et donc il en va de même pour $\sqrt{n}\left(\frac{2}{n}S_n(X) - 1\right)$, ainsi on a le test de niveau asymptotique exact suivant :

$$\psi(x) = \mathbb{1}_{\sqrt{n}\left(\frac{2}{n}S_n(x)-1\right) < -q^{\mathcal{N}(0,1)}_{1-\alpha}} = \mathbb{1}_{S_n(x) < \frac{n}{2} - \frac{\sqrt{n}}{2}q^{\mathcal{N}(0,1)}_{1-\alpha}}.$$

La méthode de Wald, précédemment utilisée pour construire des intervalles de confiance asymptoptiques, s'emploie aussi pour construire des tests asymptotiques. Il en va de même pour la méthode de stabilisation de la variance.

Exemple :

On considère un modèle d'échantillonnage avec comme statistique $\widehat{\theta}_n$ l'estimateur du maximum de vraisemblance de θ. On suppose que $\Theta \subseteq \mathbb{R}$ est un ouvert et que le modèle est régulier et que $\widehat{\theta}_n$ est consistant. Nous savons donc que

$$\forall \theta \in \Theta, \forall X \sim P_\theta,\ \sqrt{n}\left(\widehat{\theta}_n - \theta\right) \xrightarrow[n\to\infty]{\mathcal{L}} Z_\theta \sim \mathcal{N}\left(0, I(\theta)^{-1}\right) \tag{5.5}$$

La régularité du modèle implique que l'application $\theta \mapsto I(\theta)$ est continue et définie positive (ici, c'est donc un scalaire >0). La méthode de Wald pour tester $H_0 : \theta = \theta_0$ contre $H_1 : \theta \neq \theta_0$ a pour statistique de test $\widehat{\theta}_n$ région de rejet

$$\left\{\tilde{\theta} : \left|\sqrt{n}\frac{\tilde{\theta} - \theta_0}{I(\tilde{\theta})^{-1/2}}\right| > q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}\right\},$$

ou, de façon équivalente, pour statistique de test $\left|\sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{I(\widehat{\theta}_n)^{-1/2}}\right|$ et pour région de rejet $\left]q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}, \infty\right[$.

En effet, sous H_0,

$$\sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{I(\theta_0)^{-1/2}} \xrightarrow[n\to\infty]{\mathcal{L}} Z \sim \mathcal{N}(0,1),$$

par le LAC et la continuité de $x \mapsto \frac{I(\theta_0)^{-1/2}}{I(x)^{-1/2}}$, combinée à la consistance de $\widehat{\theta}_n$, $\frac{I(\theta_0)^{-1/2}}{I(\widehat{\theta}_n)^{-1/2}} \xrightarrow[n\to\infty]{\mathbb{P}} 1$ qui est une constante, donc par le corollaire multiplicatif du lemme de Slutsky,

$$\sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{I(\widehat{\theta}_n)^{-1/2}} = \frac{I(\theta_0)^{-1/2}}{I(\widehat{\theta}_n)^{-1/2}} \times \sqrt{n}\frac{\widehat{\theta}_n - \theta_0}{I(\theta_0)^{-1/2}} \xrightarrow[n\to\infty]{\mathcal{L}} 1 \times Z = Z \sim \mathcal{N}(0,1),$$

et la construction d'un test de niveau asymptotique exact α s'ensuit.

5.7 Lien avec les intervalles de confiance

On considère le test d'hypothèses $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. On suppose que l'on sait construire une région de niveau de confiance $1-\alpha$ notée $RC_{1-\alpha}(X, \theta)$ pour le paramètre θ. On peut alors construire le test de statistique de test X et de région de rejet $R(\theta_0) = \{x : \theta_0 \notin RC_{1-\alpha}(X, \theta)\}$. Il est bien de niveau α :

$$\begin{aligned}\mathbb{P}_{\theta_0}\left(X \in R(\theta_0)\right) &= \mathbb{P}_{\theta_0}\left(\theta_0 \notin RC_{1-\alpha}(X, \theta)\right) \\ &\leq \alpha.\end{aligned}$$

Réciproquement si pour tout θ_0 on dispose d'un test de niveau α pour $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ dont on note la région de rejet $\tilde{R}(\theta_0)$, on peut alors construire une région de niveau de confiance $1-\alpha$ pour θ en posant $\widetilde{RC}_{1-\alpha}(X, \theta) = \{\tilde{\theta} : X \notin \tilde{R}(\tilde{\theta})\}$. En effet

$$\begin{aligned}\mathbb{P}_{\theta_0}\left(\theta_0 \in \widetilde{RC}_{1-\alpha}(X, \theta)\right) &= \mathbb{P}_{\theta_0}\left(X \notin \tilde{R}(\theta_0)\right) \\ &= 1 - \mathbb{P}_{\theta_0}\left(X \in \tilde{R}(\theta_0)\right) \\ &\geq 1 - \alpha.\end{aligned}$$

En combinant les deux constructions on retrouve bien la région de confiance du début donc c'est consistant. En effet pour $\tilde{R} = R$:

$$\begin{aligned}\widetilde{RC}_{1-\alpha}(X, \theta) &= \{\tilde{\theta} : X \notin R(\tilde{\theta})\} \\ &= \{\tilde{\theta} : X \notin \{x : \tilde{\theta} \notin RC_{1-\alpha}(X, \theta)\}\} \\ &= \{\tilde{\theta} : X \in \{x : \tilde{\theta} \in RC_{1-\alpha}(X, \theta)\}\} \\ &= \{\tilde{\theta} : \tilde{\theta} \in RC_{1-\alpha}(X, \theta)\} \\ &= RC_{1-\alpha}(X, \theta)\end{aligned}$$

(c'est complètement tautologique).

Annexe : Lois usuelles

Nom	paramètres	$\mathbb{P}(X=k)$	$\mathbb{E}[X]$	$\mathbb{V}(X)$	Support
Bernoulli $\mathcal{B}(p)$	$p \in [0,1]$	$\mathbb{P}(X=1)=p$ $\mathbb{P}(X=0)=1-p$	p	$p(1-p)$	{0,1}
Binomiale $\mathcal{B}(n,p)$	$n \in \mathbb{N}^*$ et $p \in [0,1]$	$C_n^k p^k (1-p)^{n-k}$	np	$np(1-p)$	$[\![0,n]\!]$
Géométrique $\mathcal{G}(p)$	$p \in [0,1]$	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\mathbb{N}^*$
Poisson $\mathcal{P}(\lambda)$	$\lambda > 0$	$e^{-\lambda}\frac{\lambda^k}{k!}$	λ	λ	$\mathbb{N}$
Binomiale négative $\mathcal{NB}(n,p)$	$n \in \mathbb{N}^*$ et $p \in [0,1]$	$C_{k-1}^{n-1} p^n (1-p)^{k-n}$	$\frac{n}{p}$	$\frac{n(1-p)}{p^2}$	$[\![n,\infty[\![$
Hypergéométrique $\mathcal{H}(N,n,G)$	$N \in \mathbb{N}^*$ et $n, G \leq N$	$\frac{C_G^k C_{n-G}^{n-k}}{C_N^n}$	$n\frac{G}{N}$	$n\frac{G}{N}\frac{N-G}{N}\frac{N-n}{N-1}$	$[\![0 \wedge (n+G-N), n \vee G]\!]$

• Une v.a. de loi binomiale $\mathcal{B}(n,p)$ est la somme de n v.a. de Bernoulli indépendantes de paramètre p.

• Une v.a. X de loi binomiale négative $\mathcal{NB}(n,p)$ est la somme de n v.a. géométriques indépendantes de paramètre p. Son interprétation est donc que X est le nombre total d'essais pour accomplir n succès. Attention à cette loi qui connaît plusieurs définitions différentes, cf Wikipédia (c'est la 2e ligne du tableau de l'article qui correspond à ce poly).

• Si on tire uniformément et simultanément n individus dans une population de N individus au total, qui comporte G individus avec un trait de caractère particulier, la loi du nombre des individus tirés avec le trait de caractère est une loi hypergéométrique $\mathcal{H}(N,n,G)$.

• Les lois binomiales, de Poisson et binomiales négatives sont stables par additions indépendantes.

Nom	paramètres	densité $f(x)$	$\mathbb{E}[X]$	$\mathbb{V}(X)$	Support
Loi uniforme $\mathcal{U}([a,b])$	$(a,b) \in \mathbb{R}^2, a<b$	$\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$[a,b]$
Loi exponentielle $\mathcal{E}(\lambda)$	$\lambda > 0$	$\lambda e^{-\lambda x}\mathbb{1}_{[0,+\infty[}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$[0,+\infty[$
Loi normale $\mathcal{N}(m,\sigma^2)$	$m \in \mathbb{R}$, $\sigma^2 \geq 0$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-m)^2}{\sigma^2}\right)$ $(\sigma^2>0)$	m	σ^2	$\mathbb{R}$
Loi du chi-deux $\chi^2(n)$	$n \in \mathbb{N}^*$	$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}\exp\left(-\frac{x}{2}\right)\mathbb{1}_{[0,\infty[}(x)$	n	$2n$	$[0,+\infty[$
Loi de Student $\mathcal{T}(n)$	$n \in \mathbb{N}^*$	$\frac{1}{\sqrt{n\pi}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}$	0 $(n \geq 2)$	$\frac{n}{n-2}$ $(n \geq 3)$	$\mathbb{R}$
Loi gamma Gamma(a,λ)	$a>0, \lambda>0$	$\frac{x^{a-1}\lambda^a}{\Gamma(a)}e^{-\lambda x}\mathbb{1}_{[0,\infty[}(x)$	$\frac{a}{\lambda}$	$\frac{a}{\lambda^2}$	$[0,\infty[$
Loi beta Beta(a,b)	$a>0, b>0$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x)$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$	$[0,1]$

• On peut autoriser $\sigma^2 = 0$ dans la définition de la loi normale mais dans ce cas il n'y a pas de densité par rapport à la mesure de Lebesgue.

• La loi $\chi^2(n)$ est la loi de la somme des carrés de n lois normales centrées réduites indépendantes.

• La loi de Student $\mathcal{T}(n)$ est la loi de $\frac{U}{\sqrt{X/n}}$ avec $U \sim \mathcal{N}(0,1)$, $X \sim \chi^2(n)$, et U et X indépendantes.

• La somme de n variables exponentielles $\mathcal{E}(\lambda)$ indépendantes suit une loi Gamma $\Gamma(n,\lambda)$.

• La loi normale multi-dimensionnelle $\mathcal{N}(m,\Sigma)$, de paramètres $m \in \mathbb{R}^n$ et $\Sigma \in \mathcal{M}_{n,n}(\mathbb{R})$ symétrique et semi-définie positive, a pour espérance m, pour matrice de variance-covariance Σ, et pour support $\mathbb{R}^n$.

Si de plus Σ est définie positive (de façon équivalente, si elle est inversible) alors la loi est dominée par la mesure de Lebesgue sur $\mathbb{R}^n$ et une densité est donnée par $x \mapsto \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)\right)$.

Remerciements

Ce polycopié est tiré dans sa quasi-intégralité de sa version de 2023-2024 par Vincent Cottet, qui a lui-même tiré son matériel de Claude Petit, en charge de ce cours pour les années 2016 et 2017. Je tiens à les remercier chaleureusement tous les deux.

Chapitre 6

Travaux Dirigés : Énoncés

6.1 Probabilités

1.1 Lois discrètes usuelles

1. On dit que X suit une loi de Bernoulli de paramètre p, noté $X \sim \mathcal{B}(p)$, si $\mathcal{P}(X = 1) = p$ et $\mathcal{P}(X = 0) = 1 - p$. Calculer $\mathbb{E}[X]$ et $\mathbb{V}(X)$.

2. Soient $X_1, \ldots, X_n$ des variable indépendantes identiquement distribuées de loi $\mathcal{B}(p)$, et soit $S_n = X_1 + .. + X_n$. Quelle est la loi de S_n (le montrer par récurrence) ? On dit que S_n suit une loi binomiale de paramètres n et p, noté $S_n \sim \mathcal{B}(n, p)$. Calculer $\mathbb{E}[S_n]$ et $\mathbb{V}(S_n)$.

3. Soit N une variable aléatoire qui a pour support $\mathbb{N}$. On dit que N suit une loi de Poisson de paramètre λ, noté $N \sim \mathcal{P}(\lambda)$, s'il existe $C > 0$ tel que $\forall n \in \mathbb{N}$, $\mathbb{P}(N = n) = C\frac{\lambda^n}{n!}$. Que vaut C ? Calculer $\mathbb{E}[N]$ et $\mathbb{V}(N)$. Calculer $\mathbb{E}\left(\frac{1}{N+1}\right)$

1.2 Lois à densité usuelles

1. On dit que X suit une loi normale de paramètres m et σ, noté $X \sim \mathcal{N}(m, \sigma)$, si elle admet pour densité $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-m)^2}{2\sigma^2}}$. Calculer $\mathbb{E}[X]$ et $\mathbb{V}(X)$.

2. On dit que X suit une loi exponentielle de paramètres λ, noté $X \sim \mathcal{E}(\lambda)$, si elle admet pour densité $f(x) = Ce^{-\lambda x}\mathbb{1}_{x\geq 0}$ pour une certaine constante C. Calculer C. Quelle est sa fonction de répartition ? Calculer $\mathbb{E}[X]$ et $\mathbb{V}(X)$.

3. Soient X, Y des variables aléatoires indépendantes telles que $X \sim \mathcal{E}(\lambda)$ et $Y \sim \mathcal{E}(\lambda)$. Quelle est la loi de $X + Y$?

4. On dit que X suit une loi de Cauchy si elle admet la densité $f(x) = \frac{1}{\pi(1+x^2)}$. Une variable de Cauchy est-elle intégrable (i.e. a-t-elle une espérance) ? Quelle est la loi de $1/X$?

5. Soit U une variable aléatoire uniforme sur $[0, 1]$ (de densité $f(x) = \mathbb{1}_{0<x<1}$). Pour quelles valeurs de α la variable U^α est-elle intégrable ? Que vaut alors son espérance ? Soit $\lambda > 0$, déterminer la loi de $-\frac{1}{\lambda}\ln(U)$.

1.3 Autour de l'espérance

1. Soit X une variable aléatoire de carré intégrable à valeurs dans $\mathbb{R}$. Montrer que $\mathbb{E}\left[(a - X)^2\right]$ atteint son minimum pour $a = \mathbb{E}[X]$.

2. Soit X une variable aléatoire intégrable absolument continue de densité f et à support dans $\mathbb{R}_+$. Montrer que $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\mathrm{d}t$.

3. Soit A un évenement. Montrer que $\mathbb{E}[\mathbb{1}_A] = \mathbb{P}(A)$. En déduire que pour tout $p \geq 1$, toute variable aléatoire positive X qui appartient à L^p et tout $a > 0$, $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X^p]}{a^p}$. Montrer en particulier que pour toute variable X de carré intégrable, $\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathbb{V}(X)}{a^2}$.

1.4 Pour aller plus loin

1. Montrer qu'en fait il n'y a pas besoin de supposer que X est absolument continue dans la question 2 de l'exercice 1.3.

2. En déduire, si X est à support dans $\mathbb{N}$ et intégrable, l'égalité $\mathbb{E}[X] = \sum_{n=0}^\infty \mathbb{P}(X > n)$.

6.2 Estimation, construction d'estimateurs

2.1 Loi uniforme : différents estimateurs

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi uniforme sur $[0, \theta]$, $\theta > 0$.

1. Donner le modèle statistique associé.

2. Déterminer un estimateur $\hat{\theta}_n^{(1)}$ de θ par la méthode des moments. Étudier son biais et son risque quadratique. Montrer qu'il est consistant, de deux façons.

3. Déterminer un estimateur $\hat{\theta}_n^{(2)}$ de θ par la méthode du maximum de vraisemblance. Étudier sa loi, puis son biais et son risque quadratique. Montrer qu'il est consistant, de deux façons. En déduire un estimateur $\hat{\theta}_n^{(3)}$ de θ, sans biais.

4. Quel est le meilleur des estimateurs précédents ?

2.2 Deux estimateurs du paramètre d'une loi de Poisson

Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi de Poisson de paramètre λ, $\lambda > 0$.

1. Donner le modèle statistique associé. Est-il dominé ?

2. Déterminer deux estimateurs par la méthode des moments de λ, à partir de la moyenne et de la variance.

3. Calculer l'estimateur du maximum de vraisemblance.

4. Comparer les biais de ces estimateurs.

5. Montrer que tous ces estimateurs sont consistants.

2.3 Estimation du paramètre d'une loi de Pareto

La loi de Pareto est utilisée pour modéliser les salaires par exemple. Le paramètre est de dimension 2 et on le note : $(x_m, \lambda) \in \left(\mathbb{R}_+^*\right)^2$. C'est une loi continue dont le support est $[x_m, \infty[$ et la distribution est caractérisée par :

$$\forall x \leq x_m, \mathbb{P}(X > x) = 1,$$
$$\forall x > x_m, \mathbb{P}(X > x) = \left(\frac{x_m}{x}\right)^{\lambda}.$$

Soit $(X_1, \ldots, X_n)$ un n-échantillon de loi de Pareto $(x_m, \lambda) \in \left(\mathbb{R}_+^*\right)^2$.

1. Déterminer la vraisemblance du modèle.

2. Donner l'estimateur du maximum de vraisemblance.

3. On suppose de plus que $\lambda > 2$. Donner un estimateur des moments.

2.4 Loi binomiale négative : estimation sans biais.

Soit $(X_k)_{k \geq 1}$ une suite de v.a.i.i.d. de loi de Bernoulli de paramètre p, $p \in]0, 1[$, et soit T_n le nombre d'essais nécessaires pour obtenir n succès.

1. Déterminer la loi de T_n. Calculer son espérance et sa variance.

2. Montrer que $\widehat{p}_n = (n-1)/(T_n - 1)$ est un estimateur sans biais de p.

3. Montrer que $\mathbb{E}\left[\frac{n}{T_n}\right] > p$.

6.3 Cadre asymptotique

3.1 Loi Uniforme

On reprend le cadre de l'exercice 2.1 et les deux estimateurs.

1. Trouver la loi asymptotique de l'estimateur par la méthode des moments. Est-il asymptotiquement biaisé ?

2. Le modèle est-il régulier ?

3. Calculer la loi de $n\frac{\theta-\hat{\theta}_n^{(2)}}{\theta}$ puis étudier la convergence de celle-ci.

4. Quel est l'estimateur à privilégier dans un cadre asymptotique ?

3.2 Loi de Poisson

On reprend le cadre de l'exercice 2.2.

1. Calculer l'information de Fisher et trouver la loi limite de l'estimateur du maximum de vraisemblance de deux façons différentes.

2. Calculer la loi limite de l'estimateur des moments utilisant la variance.

3. Quel estimateur vaut-il mieux utiliser ?

3.3 Loi Gamma

La famille des lois Gamma, notée $\Gamma(\alpha, \beta)$, englobe la loi exponentielle et sert à modéliser un phénomène à support sur $\mathbb{R}_+$. La densité est donnée, pour $\alpha, \beta > 0$, par :

$$f_{\alpha,\beta}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \mathbb{1}_{\mathbb{R}^+}(x),$$

où $\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) \mathrm{d}t$ est la fonction dite *Gamma d'Euler* et a la propriété : $\Gamma(x+1) = x\Gamma(x)$. Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi Gamma de paramètres $\alpha, \beta > 0$.

1. Calculer $\Gamma(1)$ et en déduire que $\Gamma(n+1) = n!$ pour tout $n \geq 0$.

2. Donner un estimateur par la méthode des moments.

3. Donner les équations vérifiées par l'estimateur du maximum de vraisemblance.

4. Étudier la consistance de l'estimateur de la méthode des moments.

5. Donner la loi limite de l'estimateur de la méthode des moments.

6. Donner l'information de Fisher.

6.4 Intervalles de confiance

4.1 Intervalles de confiance pour une loi exponentielle

Soit $X = (X_1, \dots, X_n)$ un échantillon iid de loi $\mathcal{E}(\lambda_0)$. La loi $\mathcal{E}(\lambda)$ est continue et a pour densité $f_\lambda(x) = \lambda \exp(-\lambda x) \mathbb{1}_{\mathbb{R}^*}(x)$ pour $\lambda > 0$. La loi $\mathcal{E}(\lambda)$ est aussi la loi Gamma$(1, \lambda)$.

Une somme de variables aléatoires indépendantes de loi $\mathcal{E}(\lambda)$ suit une loi Gamma(n, λ).

1. Montrer que $\lambda_0 \sum_{i=1}^n X_i$ suit une loi Gamma$(n, 1)$. Comment appelle-t-on la variable aléatoire $Z(\lambda_0, X_1, \dots, X_n) = \lambda_0 \sum_{i=1}^n X_i$?

2. Construire un intervalle de confiance pour λ_0 non asymptotique de niveau 95%. Est-il exact ?

3. Quelle est la loi de $\lambda_0 \min_{1 \le i \le n} X_i$? Commenter.

4. Construire un autre intervalle de confiance pour λ_0 non asymptotique de niveau 95%. Est-il exact ?

5. Calculer l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n$, l'information de Fisher du modèle. Justifier la normalité asymptotique de l'EMV et donner la loi limite.

6. En déduire deux intervalles de confiance asymptotiques à 95% différents.

7. Comparer ces deux derniers intervalles, non-asymptotiquement et asymptotiquement.

4.2 Intervalles de confiance pour une loi de Bernoulli

On cherche à estimer une proportion p à partir d'un échantillon $X_1, \dots, X_n$ de v.a. de loi de Bernoulli $\mathcal{B}(p)$.

1. En utilisant l'inégalité de Bienaymé-Tchebychev, donner un intervalle de confiance pour p de niveau de confiance $1 - \alpha$ (il faudra majorer $p(1-p)$), $\alpha \in]0, 1[$.

2. Utiliser l'inégalité de Hoeffding pour donner un intervalle de confiance pour p de niveau de confiance $1 - \alpha$.

3. En utilisant le TCL, donner un intervalle de confiance asymptotique pour p de même niveau.

4.3 Intervalles de confiance pour une loi de Poisson

On reprend le cadre des exercices 2.2. et 3.2.

1. Déduire de la loi limite de l'estimateur du maximum de vraisemblance un intervalle de confiance pour λ de niveau de confiance $1 - \alpha$, $\alpha \in]0, 1[$.

2. Déduire de la loi limite de l'estimateur de la méthode des moments un intervalle de confiance pour λ de niveau de confiance $1 - \alpha$.

3. Comparer ces deux derniers intervalles asymptotiquement.