Section 1

Rappels d'inférence asymptotique et estimateurs M

Séance du vendredi 15 janvier 2027
Lectures : Wu (2010), § 9.2.2 ; complément facultatif : Boos et Stefanski (2013), ch. 7, ou Stefanski et Boos (2002)

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

À la fin de cette séance, vous devriez être en mesure de :

1. manipuler les modes de convergence (en probabilité, en loi) et les outils qui s'y rattachent (loi des grands nombres, théorème de continuité, théorème central limite multivarié, théorème de Slutsky, méthode delta d'ordre 1 et d'ordre 2) ;
2. dériver un estimateur du maximum de vraisemblance (EMV), son vecteur score et son information de Fisher, et le calculer numériquement par l'algorithme de Newton-Raphson ;
3. démontrer la normalité asymptotique de l'EMV et construire les tests de Wald, du rapport de vraisemblances et du score ;
4. définir un estimateur M à partir d'une équation d'estimation sans biais, obtenir sa loi asymptotique et sa variance « sandwich » $A^{-1}BA^{-\top}$, et l'estimer en pratique ;
5. interpréter la fonction d'influence comme le terme de linéarisation d'un estimateur asymptotiquement linéaire, outil qui reviendra aux séances 6 à 10.

Mise en contexte

Une équipe d'oncologie suit n patients après une chirurgie et note le temps (en mois) jusqu'à la récidive. Elle postule une loi de Weibull pour ces temps, estime ses deux paramètres et veut savoir si le risque de récidive est constant dans le temps (paramètre de forme égal à 1, loi exponentielle). Trois questions surgissent immédiatement : comment calculer l'estimateur quand il n'a pas de forme explicite ; comment en obtenir une erreur type ; et que deviennent ces erreurs types si la loi de Weibull n'est qu'une approximation commode de la réalité.

Les deux premières questions relèvent de la théorie classique du maximum de vraisemblance. La troisième nous amène aux *estimateurs M* : on ne suppose plus que le modèle est exact, seulement qu'une certaine équation d'estimation est centrée à la vraie valeur du paramètre d'intérêt. Cette idée

est le fil conducteur d'une bonne partie du cours : équations d'estimation généralisées (GEE, séance 6), estimation robuste, estimateurs pondérés par l'inverse de la probabilité (séance 9), échantillonnage à deux phases (séance 10). La présente séance pose les outils asymptotiques communs à toutes ces méthodes.

1 Modes de convergence et outils asymptotiques

Dans tout ce qui suit, $\|\cdot\|$ désigne une norme quelconque sur $\mathbb{R}^d$ (elles sont toutes équivalentes).

1.1 Convergence en probabilité

Définition 1.1. Une suite de vecteurs aléatoires $(X_n)_{n\geq 1}$ de $\mathbb{R}^d$ converge en probabilité vers un vecteur aléatoire X si, pour tout $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\|X_n - X\| > \epsilon) = 0.$$

On écrit $X_n \xrightarrow{\mathbb{P}} X$. Dans ce cours, la limite X sera presque toujours une constante c.

Exemple 1.2. Soit $X_n \sim \mathcal{N}(0, 1/n)$. Pour $\epsilon > 0$, $\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(|\sqrt{n}X_n| > \sqrt{n}\,\epsilon) = 2\{1 - \Phi(\sqrt{n}\,\epsilon)\} \to 0$, où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$. Donc $X_n \xrightarrow{\mathbb{P}} 0$.

Exemple 1.3. Soit $Y_1, Y_2, \ldots$ i.i.d. de loi uniforme sur $[0,1]$ et $X_n = Y_{(n)} = \max(Y_1, \ldots, Y_n)$. Pour $0 < \epsilon < 1$,

$$\mathbb{P}(|X_n - 1| > \epsilon) = \mathbb{P}(X_n < 1 - \epsilon) = \prod_{i=1}^{n} \mathbb{P}(Y_i < 1 - \epsilon) = (1-\epsilon)^n \to 0,$$

donc $Y_{(n)} \xrightarrow{\mathbb{P}} 1$.

Théorème 1.4 (Loi faible des grands nombres)**.** *Si $Y_1, Y_2, \ldots$ sont i.i.d. dans $\mathbb{R}^d$ avec $\mathrm{E}\,\|Y_1\| < \infty$ et $\mathrm{E}(Y_1) = \mu$, alors $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i \xrightarrow{\mathbb{P}} \mu$.*

Sous l'hypothèse supplémentaire $\mathrm{E}\,\|Y_1\|^2 < \infty$, la preuve tient en une ligne grâce à l'inégalité de Bienaymé-Tchebychev : $\mathbb{P}(|\bar{Y}_n - \mu| > \epsilon) \leq \mathrm{Var}(Y_1)/(n\epsilon^2) \to 0$ (cas $d = 1$). La loi *forte* des grands nombres affirme la convergence presque sûre, plus forte ; nous n'en aurons pas besoin.

Théorème 1.5 (Théorème de continuité)**.** *Si $X_n \xrightarrow{\mathbb{P}} c$ et si $h : \mathbb{R}^d \to \mathbb{R}^k$ est continue en c, alors $h(X_n) \xrightarrow{\mathbb{P}} h(c)$. Le résultat reste vrai pour une limite aléatoire X si h est continue (sur un ensemble de probabilité 1 sous la loi de X), et il vaut aussi pour la convergence en loi : $X_n \xrightarrow{\mathcal{L}} X \Rightarrow h(X_n) \xrightarrow{\mathcal{L}} h(X)$.*

Esquisse (limite constante). Par continuité en c, pour tout $\epsilon > 0$, il existe $\delta > 0$ tel que $\|x - c\| \leq \delta$ entraîne $\|h(x) - h(c)\| \leq \epsilon$. Donc $\mathbb{P}(\|h(X_n) - h(c)\| > \epsilon) \leq \mathbb{P}(\|X_n - c\| > \delta) \to 0$. □

Exemple 1.6. Si $Y_1, \ldots, Y_n$ sont i.i.d. de loi exponentielle de taux λ, la loi des grands nombres donne $\bar{Y}_n \xrightarrow{\mathbb{P}} 1/\lambda$, et le théorème de continuité avec $h(u) = 1/u$ (continue en $1/\lambda > 0$) donne $1/\bar{Y}_n \xrightarrow{\mathbb{P}} \lambda$.

1.2 Convergence en loi

Définition 1.7. La suite (X_n) converge en loi vers X si $F_{X_n}(x) \to F_X(x)$ en tout point $x \in \mathbb{R}^d$ où F_X est continue. On écrit $X_n \xrightarrow{\mathcal{L}} X$, ou encore $X_n \xrightarrow{\mathcal{L}} F_X$ (par exemple $X_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$).

Exemple 1.8 (suite de l'exemple 1.3)**.** Posons $T_n = n(1 - Y_{(n)})$. Pour $0 \leq t < n$,

$$\mathbb{P}(T_n \leq t) = \mathbb{P}\Big(Y_{(n)} \geq 1 - \tfrac{t}{n}\Big) = 1 - \Big(1 - \tfrac{t}{n}\Big)^n \longrightarrow 1 - e^{-t}.$$

Ainsi $T_n \xrightarrow{\mathcal{L}} \text{Exp}(1)$. Remarquez la vitesse : l'écart $1 - Y_{(n)}$ est d'ordre $1/n$ et non $1/\sqrt{n}$, et la loi limite n'est pas normale. Nous retrouverons ce phénomène avec l'EMV de la loi uniforme (exemple 2.3), un modèle qui viole les conditions de régularité.

Quelques liens utiles entre les deux modes de convergence :

- $X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{\mathcal{L}} X$; la réciproque est fausse en général, *sauf* si la limite est une constante : $X_n \xrightarrow{\mathcal{L}} c \iff X_n \xrightarrow{\mathbb{P}} c$.
- (Cramér-Wold) $X_n \xrightarrow{\mathcal{L}} X$ dans $\mathbb{R}^d$ si et seulement si $a^\top X_n \xrightarrow{\mathcal{L}} a^\top X$ pour tout $a \in \mathbb{R}^d$. C'est ainsi qu'on ramène le TCL multivarié au cas univarié.
- Notation stochastique : $X_n = o_{\mathbb{P}}(1)$ signifie $X_n \xrightarrow{\mathbb{P}} 0$; $X_n = O_{\mathbb{P}}(1)$ signifie que X_n est bornée en probabilité (pour tout ϵ, il existe M tel que $\mathbb{P}(\|X_n\| > M) < \epsilon$ pour tout n). Toute suite qui converge en loi est $O_{\mathbb{P}}(1)$, et $O_{\mathbb{P}}(1) \cdot o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$. Ces notations allègent considérablement les démonstrations par développement de Taylor.

Théorème 1.9 (Théorème central limite multivarié)**.** *Soit $Y_1, Y_2, \ldots$ i.i.d. dans $\mathbb{R}^d$ avec $\mathrm{E}(Y_1) = \mu$ et $\mathrm{Var}(Y_1) = \Sigma$ (matrice $d \times d$ semi-définie positive, à éléments finis). Alors*

$$\sqrt{n}\,(\bar{Y}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Sigma).$$

Théorème 1.10 (Slutsky)**.** *Si $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathbb{P}} c$ (constante), alors $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$, $Y_n X_n \xrightarrow{\mathcal{L}} cX$ et, si $c \neq 0$, $X_n / Y_n \xrightarrow{\mathcal{L}} X/c$. Dans le cas matriciel : si $M_n \xrightarrow{\mathbb{P}} M$ avec M inversible, alors $M_n^{-1} X_n \xrightarrow{\mathcal{L}} M^{-1} X$.*

Le théorème de Slutsky découle du théorème de continuité appliqué au couple (X_n, Y_n), qui converge en loi vers (X, c) précisément parce que la seconde limite est constante.

Exemple 1.11 (Statistique de Student asymptotique)**.** Soit $X_1, \ldots, X_n$ i.i.d. avec $\mathrm{E}(X_i) = \mu$ et $\mathrm{Var}(X_i) = \sigma^2 \in (0, \infty)$, et soit $S_n^2 = (n-1)^{-1} \sum_i (X_i - \bar{X}_n)^2$. On montre que $\sqrt{n}(\bar{X}_n - \mu)/S_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$.

Étape 1 : $S_n^2 \xrightarrow{\mathbb{P}} \sigma^2$. Un calcul direct donne

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

La loi des grands nombres appliquée aux $(X_i - \mu)^2$ donne $n^{-1}\sum (X_i - \mu)^2 \xrightarrow{\mathbb{P}} \sigma^2$; appliquée aux X_i, avec le théorème de continuité, elle donne $(\bar{X}_n - \mu)^2 \xrightarrow{\mathbb{P}} 0$. Enfin $n/(n-1) \to 1$.

Étape 2. Par le théorème central limite (et non la loi forte des grands nombres, comme l'indiquait par erreur la version H26 des notes), $\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$. Par continuité, $\sigma/S_n \xrightarrow{\mathbb{P}} 1$, et par Slutsky,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \cdot \frac{\sigma}{S_n} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1).$$

Aucune normalité des X_i n'est requise : c'est ce qui justifie les intervalles de type Wald $\bar{X}_n \pm 1{,}96\, S_n/\sqrt{n}$ que nous utiliserons constamment dans les études de simulation (séance 2).

1.3 Méthode delta

Théorème 1.12 (Méthode delta, ordre 1)**.** *Supposons $\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}_d(0, \Sigma)$ et soit $h : \mathbb{R}^d \to \mathbb{R}^k$ différentiable en μ, de matrice jacobienne $\nabla h(\mu)$ ($k \times d$, ligne j = gradient de la composante h_j). Alors*

$$\sqrt{n}\,\{h(X_n) - h(\mu)\} \xrightarrow{\mathcal{L}} \mathcal{N}_k\big(0, \nabla h(\mu)\,\Sigma\,\nabla h(\mu)^\top\big).$$

Pour $d = k = 1$: $\sqrt{n}\{h(X_n) - h(\mu)\} \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, h'(\mu)^2\sigma^2\big)$.

Démonstration. La différentiabilité en μ s'écrit $h(x) = h(\mu) + \nabla h(\mu)(x - \mu) + r(x)$ avec $r(x) = o(\|x - \mu\|)$. Comme $\sqrt{n}(X_n - \mu) = O_{\mathbb{P}}(1)$, on a $X_n \xrightarrow{\mathbb{P}} \mu$ et $\sqrt{n}\, r(X_n) = o_{\mathbb{P}}(1)$ (car $\sqrt{n}\, r(X_n) = \|\sqrt{n}(X_n - \mu)\| \cdot r(X_n)/\|X_n - \mu\| = O_{\mathbb{P}}(1) o_{\mathbb{P}}(1)$). Donc $\sqrt{n}\{h(X_n) - h(\mu)\} = \nabla h(\mu)\sqrt{n}(X_n - \mu) + o_{\mathbb{P}}(1)$, et la conclusion suit du théorème de continuité (application linéaire) et de Slutsky. □

Remarque 1.13. La version H26 exigeait seulement que h soit continue et affirmait que le résultat est vrai « si et seulement si » une composante de $\nabla h(\mu)$ est non nulle. Il faut plutôt : h *différentiable* en μ. Le résultat reste alors vrai même si $\nabla h(\mu) = 0$, mais la loi limite est dégénérée (une masse en 0) et ne renseigne plus sur la variabilité de $h(X_n)$; il faut alors passer à l'ordre 2.

Exemple 1.14. Avec Y_i i.i.d. exponentielles de taux λ, $\mathrm{E}(Y_i) = 1/\lambda$ et $\mathrm{Var}(Y_i) = 1/\lambda^2$. Le TCL donne $\sqrt{n}(\bar{Y}_n - 1/\lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1/\lambda^2)$. Avec $h(u) = 1/u$, $h'(1/\lambda) = -\lambda^2$, donc

$$\sqrt{n}\,(1/\bar{Y}_n - \lambda) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^4 \cdot \lambda^{-2}) = \mathcal{N}(0, \lambda^2).$$

L'estimateur $\widehat{\lambda} = 1/\bar{Y}_n$ (qui est aussi l'EMV) a donc une variance approximative λ^2/n, estimée par $\widehat{\lambda}^2/n$.

Théorème 1.15 (Méthode delta, ordre 2)**.** *Supposons $d = k = 1$, $\sqrt{n}(X_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, h deux fois différentiable en μ avec $h'(\mu) = 0$ et $h''(\mu) \neq 0$. Alors*

$$n\,\{h(X_n) - h(\mu)\} \xrightarrow{\mathcal{L}} \frac{\sigma^2 h''(\mu)}{2}\,\chi^2_1.$$

Plus généralement, en dimension d, si $\nabla h(\mu) = 0$ et H est la matrice hessienne de h en μ : $n\{h(X_n) - h(\mu)\} \xrightarrow{\mathcal{L}} \frac{1}{2} Z^\top H Z$ avec $Z \sim \mathcal{N}_d(0, \Sigma)$.

Esquisse. Le développement de Taylor à l'ordre 2 donne $h(X_n) - h(\mu) = \frac{1}{2}h''(\mu)(X_n - \mu)^2 + o_{\mathbb{P}}(n^{-1})$, le terme d'ordre 1 étant nul. Donc $n\{h(X_n) - h(\mu)\} = \frac{1}{2}h''(\mu)\{\sqrt{n}(X_n - \mu)\}^2 + o_{\mathbb{P}}(1)$. Or $\sqrt{n}(X_n - \mu)/\sigma \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,1)$ et, par continuité, $\{\sqrt{n}(X_n - \mu)\}^2 \xrightarrow{\mathcal{L}} \sigma^2 Z^2 = \sigma^2 \chi^2_1$. □

Deux observations importantes :

— la vitesse change : l'écart $h(X_n) - h(\mu)$ est d'ordre $1/n$ et non $1/\sqrt{n}$;

— la loi limite n'est plus centrée : $\mathrm{E}\{\frac{1}{2}\sigma^2 h''(\mu)\chi_1^2\} = \frac{1}{2}\sigma^2 h''(\mu)$ et sa variance vaut $\frac{1}{2}\sigma^4 h''(\mu)^2$ (car $\mathrm{Var}(\chi_1^2) = 2$).

Exemple 1.16. Soit X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$ et $h(u) = u^2$. Si $\mu \neq 0$, l'ordre 1 donne $\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 4\mu^2\sigma^2)$. Si $\mu = 0$, $h'(0) = 0$ et $h''(0) = 2$, et l'ordre 2 donne $n\bar{X}_n^2 \xrightarrow{\mathcal{L}} \sigma^2\chi_1^2$ (ce qui est d'ailleurs exact pour tout n dans ce cas gaussien). La variance « d'ordre 1 », $4\mu^2\sigma^2/n$, tend vers 0 quand $\mu \to 0$ alors que la vraie variance de $\bar{X}_n^2$ ne s'annule pas : près d'un point où la dérivée s'annule, l'approximation d'ordre 1 peut être très mauvaise à taille d'échantillon finie, et le terme d'ordre 2 doit être pris en compte.

À retenir

— Loi des grands nombres + continuité ⇒ convergence ; TCL + Slutsky ⇒ loi asymptotique normale des statistiques « studentisées ».

— Méthode delta d'ordre 1 : variance asymptotique $\nabla h \, \Sigma \, \nabla h^\top$; exige h différentiable.

— Si $\nabla h(\mu) = 0$: vitesse n, loi limite $\frac{1}{2}\sigma^2 h''(\mu)\chi_1^2$ (ordre 2).

— Les vitesses non standard (ordre n) apparaissent aussi quand le support dépend du paramètre.

2 Estimation par maximum de vraisemblance

2.1 Notation et conditions de régularité

Soit $X_1, \ldots, X_n$ i.i.d. de densité (ou fonction de masse) f_θ, $\theta \in \Theta \subset \mathbb{R}^p$, et θ_0 la vraie valeur. On note

$$L(\theta) = \prod_{i=1}^n f_\theta(X_i), \qquad \ell(\theta) = \log L(\theta) = \sum_{i=1}^n \ell_i(\theta), \quad \ell_i(\theta) = \log f_\theta(X_i),$$

et l'EMV $\widehat{\theta} = \arg\max_{\theta\in\Theta} \ell(\theta)$. Les résultats asymptotiques qui suivent reposent sur des *conditions de régularité*, dont voici une version suffisante pour ce cours :

(R1) *Identifiabilité* : $\theta \neq \theta' \Rightarrow f_\theta \neq f_{\theta'}$.

(R2) *Support fixe* : $\{x : f_\theta(x) > 0\}$ ne dépend pas de θ.

(R3) *Intérieur* : θ_0 est un point intérieur de Θ (en H26 : Θ compact contenant θ_0 en son intérieur).

(R4) *Dérivabilité* : $\theta \mapsto f_\theta(x)$ est trois fois continûment différentiable, et les dérivées troisièmes de $\log f_\theta(x)$ sont dominées, au voisinage de θ_0, par une fonction $M(x)$ d'espérance finie.

(R5) *Permutation* : on peut dériver deux fois $\int f_\theta(x)\,\mathrm{d}x = 1$ sous le signe intégral.

(R6) *Information* : la matrice d'information $I(\theta_0)$ définie plus bas est finie et définie positive.

Exemple 2.1 (Loi de Bernoulli)**.** Si $X_i \in \{0,1\}$ avec $f_\theta(x) = \theta^x(1-\theta)^{1-x}$, $\theta \in (0,1)$, le support $\{0,1\}$ ne dépend pas de θ, f_θ est indéfiniment dérivable et $(0,1)$ est ouvert. Avec $Y = \sum_i X_i$,

$$\ell(\theta) = Y \log\theta + (n-Y)\log(1-\theta), \qquad \ell'(\theta) = \frac{Y}{\theta} - \frac{n-Y}{1-\theta},$$

et $\ell'(\widehat{\theta}) = 0$ donne $\widehat{\theta} = Y/n = \bar{X}_n$ (un maximum, car $\ell'' < 0$), pourvu que $0 < Y < n$.

Exemple 2.2 (Loi normale bivariée)**.** Soit $X_i = (X_{i1}, X_{i2})^\top \sim \mathcal{N}_2(\mu, \Sigma)$ avec $\mu = (\mu_1, \mu_2)^\top$ et $\Sigma = \left(\begin{smallmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{smallmatrix}\right)$; ici $d = 2$ et $p = 5$, $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})^\top$. La log-vraisemblance s'écrit

$$\ell(\mu, \Sigma) = -n\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^\top \Sigma^{-1}(X_i - \mu).$$

Pour Σ fixée, la somme quadratique est minimisée en $\mu = \bar{X}$ (décomposer $X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$). En posant $W = \sum_i (X_i - \bar{X})(X_i - \bar{X})^\top$ et $\Omega = \Sigma^{-1}$, la log-vraisemblance profilée vaut, à une constante près, $\frac{n}{2}\log|\Omega| - \frac{1}{2}\operatorname{tr}(\Omega W)$. Avec $\partial \log|\Omega|/\partial\Omega = \Omega^{-1}$ et $\partial \operatorname{tr}(\Omega W)/\partial\Omega = W$, on obtient $\widehat{\Sigma} = W/n$, c'est-à-dire

$$\widehat{\mu}_j = \frac{1}{n}\sum_i X_{ij}, \qquad \widehat{\sigma}_j^2 = \frac{1}{n}\sum_i (X_{ij} - \widehat{\mu}_j)^2, \qquad \widehat{\sigma}_{12} = \frac{1}{n}\sum_i (X_{i1} - \widehat{\mu}_1)(X_{i2} - \widehat{\mu}_2).$$

Notez le diviseur n (et non $n-1$) : l'EMV de la variance est biaisé, mais asymptotiquement sans biais.

Exemple 2.3 (Loi uniforme : un cas non régulier)**.** Soit X_i i.i.d. uniformes sur $[0, \theta]$, $f_\theta(x) = \theta^{-1}\mathbf{1}\{0 \le x \le \theta\}$. Le support dépend de θ : (R2) est violée. La vraisemblance

$$L(\theta) = \theta^{-n}\,\mathbf{1}\{\theta \ge X_{(n)}\}$$

est nulle pour $\theta < X_{(n)}$ et décroissante ensuite : $\widehat{\theta} = X_{(n)}$. L'équation $\ell'(\theta) = 0$ n'a pas de solution, et la théorie qui suit ne s'applique pas. D'après l'exemple 1.8 (appliqué à X_i/θ), $n(\theta - \widehat{\theta})/\theta \xrightarrow{\mathcal{L}} \text{Exp}(1)$: convergence à vitesse n vers une loi non normale.

2.2 Vecteur score et information de Fisher

Sous les conditions de régularité, on définit le vecteur score (de longueur p) et la matrice des dérivées secondes :

$$S(\theta) = \frac{\partial \ell(\theta)}{\partial\theta} = \sum_{i=1}^{n} S_i(\theta), \qquad S_i(\theta) = \frac{\partial \log f_\theta(X_i)}{\partial\theta},$$

$$U(\theta) = \frac{\partial S(\theta)}{\partial\theta^\top} = \sum_{i=1}^{n} U_i(\theta), \qquad U_i(\theta) = \frac{\partial^2 \log f_\theta(X_i)}{\partial\theta\,\partial\theta^\top}.$$

L'EMV (intérieur) est solution de l'équation de vraisemblance $S(\widehat{\theta}) = 0$.

Proposition 2.4 (Identités de Bartlett)**.** *Sous (R2), (R4) et (R5), pour tout i et en notant E_θ l'espérance sous f_θ :*

(a) $\mathrm{E}_\theta\{S_i(\theta)\} = 0$; *(b)* $\mathrm{Var}_\theta\{S_i(\theta)\} = \mathrm{E}_\theta\{S_i(\theta)S_i(\theta)^\top\} = -\mathrm{E}_\theta\{U_i(\theta)\} =: I(\theta)$.

*La matrice $I(\theta)$ est l'*information de Fisher *d'une observation.*

Démonstration. (a) Comme $\partial \log f_\theta / \partial\theta = (\partial f_\theta/\partial\theta)/f_\theta$,

$$\mathrm{E}_\theta\{S_i(\theta)\} = \int \frac{\partial f_\theta(x)}{\partial\theta}\,\frac{1}{f_\theta(x)}\,f_\theta(x)\,\mathrm{d}x = \int \frac{\partial f_\theta(x)}{\partial\theta}\,\mathrm{d}x = \frac{\partial}{\partial\theta}\int f_\theta(x)\,\mathrm{d}x = \frac{\partial 1}{\partial\theta} = 0,$$

la permutation dérivée-intégrale étant permise par (R5) ; c'est ici que la condition de support fixe (R2) est cruciale. (b) Dérivons l'identité $\int\{\partial \log f_\theta(x)/\partial\theta\} f_\theta(x)\,\mathrm{d}x = 0$ par rapport à $\theta^\top$:

$$0 = \int \frac{\partial^2 \log f_\theta(x)}{\partial\theta\,\partial\theta^\top} f_\theta(x)\,\mathrm{d}x + \int \frac{\partial \log f_\theta(x)}{\partial\theta}\,\frac{\partial f_\theta(x)}{\partial\theta^\top}\,\mathrm{d}x.$$

Dans la seconde intégrale, on écrit $\partial f_\theta/\partial\theta^\top = f_\theta\,\partial \log f_\theta/\partial\theta^\top$: elle devient $\mathrm{E}_\theta\{S_i S_i^\top\}$. La première vaut $\mathrm{E}_\theta(U_i)$. Donc $\mathrm{E}_\theta(S_i S_i^\top) = -\,\mathrm{E}_\theta(U_i)$, et comme $\mathrm{E}_\theta(S_i) = 0$, $\mathrm{E}_\theta(S_i S_i^\top) = \mathrm{Var}_\theta(S_i)$. □

Par indépendance, $\mathrm{E}\{S(\theta_0)\} = 0$ et $\mathrm{Var}\{S(\theta_0)\} = nI(\theta_0)$, et le TCL donne

$$\frac{1}{\sqrt{n}} S(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, I(\theta_0)). \tag{1}$$

En pratique, on distingue l'information *attendue* $nI(\widehat\theta)$ et l'information *observée* $-U(\widehat\theta)$; les deux sont des estimateurs convergents de $nI(\theta_0)$ (après division par n), et les logiciels utilisent souvent la seconde.

Exemple 2.5 (Loi normale)**.** Pour $X_i \sim \mathcal{N}(\mu, \sigma^2)$ et $\theta = (\mu, \sigma^2)^\top$, $\log f_\theta(x) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - (x-\mu)^2/(2\sigma^2)$, d'où

$$S_i(\theta) = \begin{pmatrix} (X_i-\mu)/\sigma^2 \\ -1/(2\sigma^2) + (X_i-\mu)^2/(2\sigma^4) \end{pmatrix}, \qquad U_i(\theta) = \begin{pmatrix} -1/\sigma^2 & -(X_i-\mu)/\sigma^4 \\ -(X_i-\mu)/\sigma^4 & 1/(2\sigma^4) - (X_i-\mu)^2/\sigma^6 \end{pmatrix}.$$

On a bien $\mathrm{E}(S_i) = 0$, et $-\,\mathrm{E}(U_i) = \mathrm{diag}\big(1/\sigma^2, 1/(2\sigma^4)\big) = I(\theta)$. On vérifie l'identité (b) directement à l'aide de $\mathrm{E}(X_i-\mu)^3 = 0$ et $\mathrm{E}(X_i-\mu)^4 = 3\sigma^4$: par exemple $\mathrm{Var}\{(X_i-\mu)^2/(2\sigma^4)\} = (3\sigma^4-\sigma^4)/(4\sigma^8) = 1/(2\sigma^4)$.

2.3 Calcul numérique : l'algorithme de Newton-Raphson

Dans la plupart des modèles du cours (Weibull, régression logistique, modèles mixtes), l'équation $S(\theta) = 0$ n'a pas de solution explicite. L'algorithme de Newton-Raphson linéarise le score autour de l'itéré courant :

$$0 = S(\widehat\theta) \approx S(\widehat\theta^{(k)}) + U(\widehat\theta^{(k)})\,(\widehat\theta - \widehat\theta^{(k)}),$$

et résout l'approximation linéaire :

$$\boxed{\widehat\theta^{(k+1)} = \widehat\theta^{(k)} - \big\{U(\widehat\theta^{(k)})\big\}^{-1} S(\widehat\theta^{(k)})}$$

On arrête lorsque $\|\widehat\theta^{(k+1)} - \widehat\theta^{(k)}\| < \epsilon$ (ou lorsque $\|S\|$ est petit). Près de la solution, la convergence est quadratique : le nombre de décimales exactes double à chaque itération. Loin de la solution, l'algorithme peut diverger ; on le stabilise en réduisant le pas (recherche linéaire) ou en remplaçant $-U$ par l'information attendue $nI(\theta)$ (*méthode du score de Fisher*, utilisée par `glm`).

Exemple 2.6 (Loi de Weibull)**.** Pour la loi de Weibull de forme α et d'échelle λ, $f_\theta(x) = (\alpha/\lambda)(x/\lambda)^{\alpha-1}\exp\{-(x/\lambda)^\alpha\}$, $x > 0$. Avec $z_i = X_i/\lambda$,

$$\ell_i(\theta) = \log\alpha - \log\lambda + (\alpha-1)\log z_i - z_i^\alpha, \qquad S_i(\theta) = \begin{pmatrix} 1/\alpha + \log z_i - z_i^\alpha \log z_i \\ -\alpha/\lambda + \alpha z_i^\alpha/\lambda \end{pmatrix}.$$

Il n'y a pas de forme explicite pour $\widehat{\theta}$. Le code suivant simule 80 temps de récidive (en mois) et programme Newton-Raphson à partir de la loi exponentielle ($\alpha = 1$, $\lambda = \bar{X}$) comme valeur initiale.

```
set.seed(7360)
n <- 80
x <- rweibull(n, shape = 1.5, scale = 24)   # temps de récidive (mois)
score <- function(th, x) {
  a <- th[1]; l <- th[2]; z <- x / l
  c(sum(1/a + log(z) - z^a * log(z)), sum(-a/l + a * z^a / l))
}
hess <- function(th, x) {
  a <- th[1]; l <- th[2]; z <- x / l; lz <- log(z)
  h11 <- sum(-1/a^2 - z^a * lz^2)
  h12 <- sum(-1/l + z^a / l + a * z^a * lz / l)
  h22 <- sum(a/l^2 - a * (a + 1) * z^a / l^2)
  matrix(c(h11, h12, h12, h22), 2, 2)
}
th <- c(1, mean(x))                         # départ : loi exponentielle
for (k in 1:20) {
  pas <- solve(hess(th, x), score(th, x))
  th <- th - pas
  cat(sprintf("k = %2d  alpha = %.6f  lambda = %.6f  |pas| = %.2e\n",
              k, th[1], th[2], sqrt(sum(pas^2))))
  if (sqrt(sum(pas^2)) < 1e-8) break
}
V <- solve(-hess(th, x))                    # inverse de l'information observée
round(cbind(estimation = th, erreur.type = sqrt(diag(V))), 4)
MASS::fitdistr(x, "weibull")                # vérification
```

```
k =  1  alpha = 1.462199  lambda = 28.074672  |pas| = 2.23e+00
k =  2  alpha = 1.699447  lambda = 28.961608  |pas| = 9.18e-01
k =  3  alpha = 1.734815  lambda = 29.105113  |pas| = 1.48e-01
k =  4  alpha = 1.735427  lambda = 29.107001  |pas| = 1.98e-03
k =  5  alpha = 1.735427  lambda = 29.107002  |pas| = 8.98e-07
k =  6  alpha = 1.735427  lambda = 29.107002  |pas| = 2.65e-14
     estimation erreur.type
[1,]     1.7354      0.1549
[2,]    29.1070      1.9759
     shape        scale
   1.7356742   29.1070473
 ( 0.1548761) ( 1.9756004)
```

La convergence quadratique est visible dans la dernière colonne (10^{-3}, 10^{-7}, 10^{-14}). Les petites différences avec `fitdistr` (quatrième décimale) viennent du critère d'arrêt de l'optimiseur générique `optim` qu'il emploie. Le test de $H_0 : \alpha = 1$ (risque constant) est l'objet de la section 3.

2.4 Propriétés asymptotiques de l'EMV

Théorème 2.7 (Convergence)**.** *Sous (R1)–(R4), il existe une suite de racines de l'équation de vraisemblance telle que* $\widehat{\theta} \xrightarrow{\mathbb{P}} \theta_0$.

Idée. Par la loi des grands nombres, $n^{-1}\ell(\theta) \xrightarrow{\mathbb{P}} \mathrm{E}_{\theta_0}\{\log f_\theta(X)\}$ pour chaque θ. Or, par l'inégalité de Jensen,

$$\mathrm{E}_{\theta_0}\Big\{\log \frac{f_\theta(X)}{f_{\theta_0}(X)}\Big\} \leq \log \mathrm{E}_{\theta_0}\Big\{\frac{f_\theta(X)}{f_{\theta_0}(X)}\Big\} = \log 1 = 0,$$

avec égalité seulement si $f_\theta = f_{\theta_0}$, donc (R1) seulement si $\theta = \theta_0$. La limite de $n^{-1}\ell$ est donc maximisée uniquement en θ_0 : la quantité $\mathrm{KL}(\theta_0, \theta) = \mathrm{E}_{\theta_0} \log\{f_{\theta_0}(X)/f_\theta(X)\} \geq 0$ est la divergence de Kullback-Leibler. Une convergence uniforme (ou un argument local) permet de passer des maximiseurs des fonctions aux maximiseurs des limites. □

Si le modèle est *mal spécifié* (les données suivent une loi G qui n'appartient pas à $\{f_\theta\}$), le même argument montre que $\widehat{\theta}$ converge vers le *pseudo-vrai paramètre* $\theta^* = \arg\min_\theta \mathrm{E}_G \log\{g(X)/f_\theta(X)\}$, la loi du modèle la plus proche de G au sens de Kullback-Leibler. Nous y reviendrons à la section 4.

Théorème 2.8 (Normalité asymptotique de l'EMV)**.** *Sous (R1)–(R6), si* $\widehat{\theta}$ *est une racine convergente de* $S(\theta) = 0$*,*

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, I(\theta_0)^{-1}).$$

Démonstration. Développons chaque composante du score autour de θ_0 (théorème de la valeur moyenne) :

$$0 = S(\widehat{\theta}) = S(\theta_0) + U(\theta^*)\,(\widehat{\theta} - \theta_0),$$

où chaque ligne de $U(\theta^*)$ est évaluée en un point θ^* (propre à la ligne) du segment $[\theta_0, \widehat{\theta}]$. En multipliant par $1/\sqrt{n}$ et en réarrangeant,

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) = \Big\{-\frac{1}{n}U(\theta^*)\Big\}^{-1} \frac{1}{\sqrt{n}} S(\theta_0).$$

Terme de droite. Par (1), $n^{-1/2}S(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, I(\theta_0))$.

Matrice. Par la loi des grands nombres, $-n^{-1}U(\theta_0) = -n^{-1}\sum_i U_i(\theta_0) \xrightarrow{\mathbb{P}} -\mathrm{E}\{U_i(\theta_0)\} = I(\theta_0)$. De plus, (R4) donne, élément par élément, $\|n^{-1}U(\theta^*) - n^{-1}U(\theta_0)\| \leq \|\theta^* - \theta_0\| \cdot n^{-1}\sum_i M(X_i) = o_{\mathbb{P}}(1) \cdot O_{\mathbb{P}}(1)$, puisque $\theta^* \xrightarrow{\mathbb{P}} \theta_0$ (il est coincé entre θ_0 et $\widehat{\theta}$). Donc $-n^{-1}U(\theta^*) \xrightarrow{\mathbb{P}} I(\theta_0)$, inversible par (R6).

Conclusion. Par Slutsky (version matricielle), $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} I(\theta_0)^{-1}Z$ avec $Z \sim \mathcal{N}_p(0, I(\theta_0))$, c'est-à-dire une loi $\mathcal{N}_p(0, I(\theta_0)^{-1}I(\theta_0)I(\theta_0)^{-1}) = \mathcal{N}_p(0, I(\theta_0)^{-1})$. □

Gardez en tête la forme $I^{-1}\,I\,I^{-1}$ de la dernière ligne : le « pain » I^{-1} vient de la dérivée du score, la « garniture » I de la variance du score. Ils coïncident grâce à l'identité de Bartlett (b), qui n'est vraie que si le modèle est correct. C'est exactement ce qui distinguera les estimateurs M.

Exemple 2.9 (suite de l'exemple 2.5)**.** La résolution de $S(\theta) = 0$ donne $\widehat{\mu} = \bar{X}$ et $\widehat{\sigma}^2 = n^{-1}\sum(X_i - \bar{X})^2$. Comme $I(\theta)^{-1} = \mathrm{diag}(\sigma^2, 2\sigma^4)$, on a asymptotiquement $n\,\mathrm{Var}(\widehat{\mu}) \to \sigma^2$, $n\,\mathrm{Var}(\widehat{\sigma}^2) \to 2\sigma^4$ et $n\,\mathrm{Cov}(\widehat{\mu}, \widehat{\sigma}^2) \to 0$.

Remarque 2.10 (Efficacité). Sous les mêmes conditions, tout estimateur asymptotiquement sans biais « régulier » a une variance asymptotique au moins égale à $I(\theta_0)^{-1}$ (borne de Cramér-Rao, au sens matriciel). L'EMV est donc asymptotiquement efficace *si le modèle est correct*. L'exercice 5.3 illustre la perte d'efficacité de l'estimateur des moments.

Grille méthodologique : maximum de vraisemblance

1. **Objet d'intérêt** : le paramètre θ d'un modèle paramétrique $\{f_\theta\}$ supposé correct.
2. **Nuisance** : composantes de θ non visées par la question (p. ex. l'échelle λ quand on teste la forme α) ; biais si le modèle est faux.
3. **Traitement de la nuisance** : estimation conjointe, puis profilage (tests du RV, du score).
4. **Critère** : maximisation de $\ell(\theta)$, c'est-à-dire résolution de $S(\theta) = 0$.
5. **Calcul** : Newton-Raphson, score de Fisher, ou optimiseur générique (`optim`).
6. **Inférence** : $\widehat{\theta} \approx \mathcal{N}(\theta_0, \{nI(\widehat{\theta})\}^{-1})$ ou $\{-U(\widehat{\theta})\}^{-1}$; tests de Wald, du RV, du score.

3 Tests d'hypothèses asymptotiques

On partitionne $\theta = (\theta_1^\top, \theta_2^\top)^\top$ avec $\theta_1 \in \mathbb{R}^{p_1}$ (paramètre d'intérêt) et $\theta_2 \in \mathbb{R}^{p_2}$ (nuisance), $p_1 + p_2 = p$, et on teste

$$H_0 : \theta_1 = \theta_{10} \qquad \text{contre} \qquad H_1 : \theta_1 \neq \theta_{10}.$$

On note $\widehat{\theta} = (\widehat{\theta}_1, \widehat{\theta}_2)$ l'EMV non contraint et $\widetilde{\theta} = (\theta_{10}, \widehat{\theta}_{20})$ l'EMV sous H_0, où $\widehat{\theta}_{20}$ maximise $\ell(\theta_{10}, \theta_2)$ en θ_2. On partitionne conformément

$$I(\theta) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \qquad I(\theta)^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}, \qquad I^{11} = (I_{11} - I_{12} I_{22}^{-1} I_{21})^{-1}.$$

Par le théorème 2.8, $\sqrt{n}(\widehat{\theta}_1 - \theta_1) \xrightarrow{\mathcal{L}} \mathcal{N}_{p_1}(0, I^{11})$: la variance asymptotique de $\widehat{\theta}_1$ est le bloc $(1,1)$ de *l'inverse* de l'information, qui tient compte de l'estimation de la nuisance θ_2.

3.1 Test de Wald

$$Q_{\mathrm{W}} = (\widehat{\theta}_1 - \theta_{10})^\top \, \widehat{\Sigma}_{11}^{-1} \, (\widehat{\theta}_1 - \theta_{10}), \qquad \widehat{\Sigma}_{11} = \text{bloc } (1,1) \text{ de } \{nI(\widehat{\theta})\}^{-1} \text{ (ou de } \{-U(\widehat{\theta})\}^{-1}).$$

Sous H_0, $Q_{\mathrm{W}} \xrightarrow{\mathcal{L}} \chi^2_{p_1}$: en effet $n^{1/2}(\widehat{\theta}_1 - \theta_{10}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{11})$, et $n\widehat{\Sigma}_{11} \xrightarrow{\mathbb{P}} I^{11}$ par convergence de $\widehat{\theta}$ et continuité de $I(\cdot)$; on conclut par Slutsky et le théorème de continuité, puisque $Z^\top \Sigma^{-1} Z \sim \chi^2_{p_1}$ si $Z \sim \mathcal{N}_{p_1}(0, \Sigma)$. On rejette H_0 au niveau α si $Q_{\mathrm{W}} > \chi^2_{p_1, 1-\alpha}$ (quantile d'ordre $1-\alpha$). Si $p_1 = 1$, on utilise de façon équivalente $R_{\mathrm{W}} = (\widehat{\theta}_1 - \theta_{10}) / \widehat{\Sigma}_{11}^{1/2} \approx \mathcal{N}(0,1)$.

Correction du document H26

Dans les notes H26, Σ_{11} est défini comme le bloc de $\{nI(\theta)\}^{-1}$, puis la statistique univariée est écrite $(\widehat{\theta}_1 - \theta_{10}) / \sqrt{\Sigma_{11}/n}$: le facteur n est compté deux fois. Avec Σ_{11} tiré de $\{nI\}^{-1}$, c'est $\sqrt{\Sigma_{11}}$ qui est l'erreur type. Le théorème qui permet de remplacer $I(\theta)$ par $I(\widehat{\theta})$ sans changer

la loi limite est le théorème de Slutsky (avec le théorème de continuité et la convergence de $\widehat{\theta}$).

3.2 Test du rapport de vraisemblances

$$Q_{\mathrm{RV}} = -2\{\ell(\widetilde{\theta}) - \ell(\widehat{\theta})\} = 2\{\ell(\widehat{\theta}_1, \widehat{\theta}_2) - \ell(\theta_{10}, \widehat{\theta}_{20})\} \xrightarrow{\mathcal{L}} \chi^2_{p_1} \quad \text{sous } H_0.$$

Esquisse dans le cas simple $p_1 = p$. Développons ℓ autour de $\widehat{\theta}$ à l'ordre 2, en utilisant $S(\widehat{\theta}) = 0$: $\ell(\theta_0) = \ell(\widehat{\theta}) + \frac{1}{2}(\theta_0 - \widehat{\theta})^\top U(\theta^*)(\theta_0 - \widehat{\theta})$. Donc

$$Q_{\mathrm{RV}} = \sqrt{n}(\widehat{\theta} - \theta_0)^\top \Big\{-\tfrac{1}{n}U(\theta^*)\Big\}\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} Z^\top I(\theta_0) Z, \qquad Z \sim \mathcal{N}_p(0, I(\theta_0)^{-1}),$$

qui suit une loi χ^2_p. Le cas composite ($p_1 < p$) se traite de la même façon en développant aussi $\ell(\theta_{10}, \widehat{\theta}_{20})$; on retrouve une forme quadratique en $\widehat{\theta}_1 - \theta_{10}$ de matrice $(I^{11})^{-1}$. □

3.3 Test du score

Le test du score n'utilise que l'ajustement sous H_0 :

$$Q_{\mathrm{S}} = S(\widetilde{\theta})^\top \{nI(\widetilde{\theta})\}^{-1} S(\widetilde{\theta}) \xrightarrow{\mathcal{L}} \chi^2_{p_1} \quad \text{sous } H_0.$$

Comme la composante $S_2(\widetilde{\theta})$ du score est nulle par définition de $\widehat{\theta}_{20}$, on a aussi

$$Q_{\mathrm{S}} = n^{-1} S_1(\widetilde{\theta})^\top I^{11}(\widetilde{\theta})\, S_1(\widetilde{\theta}).$$

L'intuition : sous H_0, le score évalué en $\widetilde{\theta}$ devrait être proche de 0 ; on mesure son écart dans la métrique de sa propre variance.

Correction du document H26

La version H26 écrivait $Q_{\mathrm{S}} = S^\top [nI]\, S$. Il faut l'*inverse* de l'information : $Q_{\mathrm{S}} = S^\top [nI]^{-1} S$. Vérifiez la cohérence des ordres de grandeur : $S = O_{\mathbb{P}}(\sqrt{n})$ et $nI = O(n)$, donc $S^\top (nI)^{-1} S = O_{\mathbb{P}}(1)$, alors que $S^\top (nI) S$ croîtrait comme n^2.

3.4 Comparaison et exemple numérique

Les trois statistiques sont asymptotiquement équivalentes sous H_0 (et sous des alternatives locales), mais diffèrent à taille finie. La figure 1 (gauche) en donne la géométrie pour $p = 1$: Wald mesure la distance horizontale $\widehat{\theta} - \theta_0$, le RV la distance verticale $\ell(\widehat{\theta}) - \ell(\theta_0)$, et le score la pente de ℓ en θ_0. Quelques propriétés pratiques :

— Wald n'exige que l'ajustement non contraint, mais n'est pas invariant par reparamétrisation (tester $\theta = \theta_0$ ou $\log\theta = \log\theta_0$ ne donne pas la même statistique) ; il se comporte mal près de la frontière de Θ.

— Le RV est invariant par reparamétrisation et généralement le plus fiable ; il exige les deux ajustements.

— Le score n'exige que l'ajustement sous H_0 (pratique quand le modèle complet est difficile à ajuster) ; c'est la base de nombreux tests classiques (khi-deux de Pearson, test du log-rang).

Exemple 3.1. Dans un essai de phase II, 31 patients sur $n = 120$ présentent un effet indésirable de grade 3 ; le taux historique est de 18 %. On teste $H_0 : \theta = 0{,}18$ dans le modèle binomial. Avec $\widehat{\theta} = 31/120 = 0{,}2583$:

$$Q_{\mathrm{W}} = \frac{(\widehat{\theta} - \theta_0)^2}{\widehat{\theta}(1-\widehat{\theta})/n}, \qquad Q_{\mathrm{RV}} = 2\{\ell(\widehat{\theta}) - \ell(\theta_0)\}, \qquad Q_{\mathrm{S}} = \frac{(Y - n\theta_0)^2}{n\theta_0(1-\theta_0)}.$$

```
y <- 31; n <- 120; p0 <- 0.18; p <- y/n
W  <- (p - p0)^2 / (p*(1-p)/n)
ll <- function(t) y*log(t) + (n-y)*log(1-t)
LR <- 2*(ll(p) - ll(p0))
S  <- (y - n*p0)^2 / (n*p0*(1-p0))
round(c(phat = p, Wald = W, RV = LR, Score = S), 4)
print(1 - pchisq(c(W, LR, S), 1), digits = 4)
prop.test(y, n, p = p0, correct = FALSE)$statistic   # = test du score
```

```
  phat   Wald     RV  Score
0.2583 3.8431 4.5282 4.9887
[1] 0.04995 0.03334 0.02551
X-squared
 4.988708
```

Les trois tests rejettent H_0 à 5 %, mais de justesse pour Wald. Ici $\widehat{\theta} > \theta_0$ et la variance $\widehat{\theta}(1-\widehat{\theta})$ utilisée par Wald est plus grande que $\theta_0(1-\theta_0)$ utilisée par le score, d'où $Q_{\mathrm{W}} < Q_{\mathrm{S}}$. Le test khi-deux de `prop.test` sans correction de continuité est exactement le test du score.

À retenir

— Score centré et de variance $nI(\theta)$; $I = -\,\mathrm{E}(U) = \mathrm{E}(SS^\top)$ si le modèle est correct.

— $\widehat{\theta} \approx \mathcal{N}(\theta_0, \{nI(\theta_0)\}^{-1})$: démonstration par Taylor du score + TCL + LGN + Slutsky.

— Wald, RV et score : trois statistiques $\chi^2_{p_1}$ asymptotiquement équivalentes ; la variance de $\widehat{\theta}_1$ est le bloc de l'*inverse* de l'information, et le score utilise $\{nI\}^{-1}$.

4 Estimateurs M

4.1 Définition

L'EMV résout $\sum_i S_i(\theta) = 0$. Rien n'oblige la fonction sommée à être un score : il suffit qu'elle soit centrée à la bonne valeur du paramètre.

Définition 4.1. Soit $Y_1, \ldots, Y_n$ i.i.d. de loi F (non nécessairement paramétrique) et $\psi(y, \theta) \in \mathbb{R}^p$ une fonction, $\theta \in \mathbb{R}^p$. Un *estimateur M* (ou estimateur Z) $\widehat{\theta}$ est une solution de l'*équation d'estimation*

$$\Psi_n(\widehat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i, \widehat{\theta}) = 0.$$

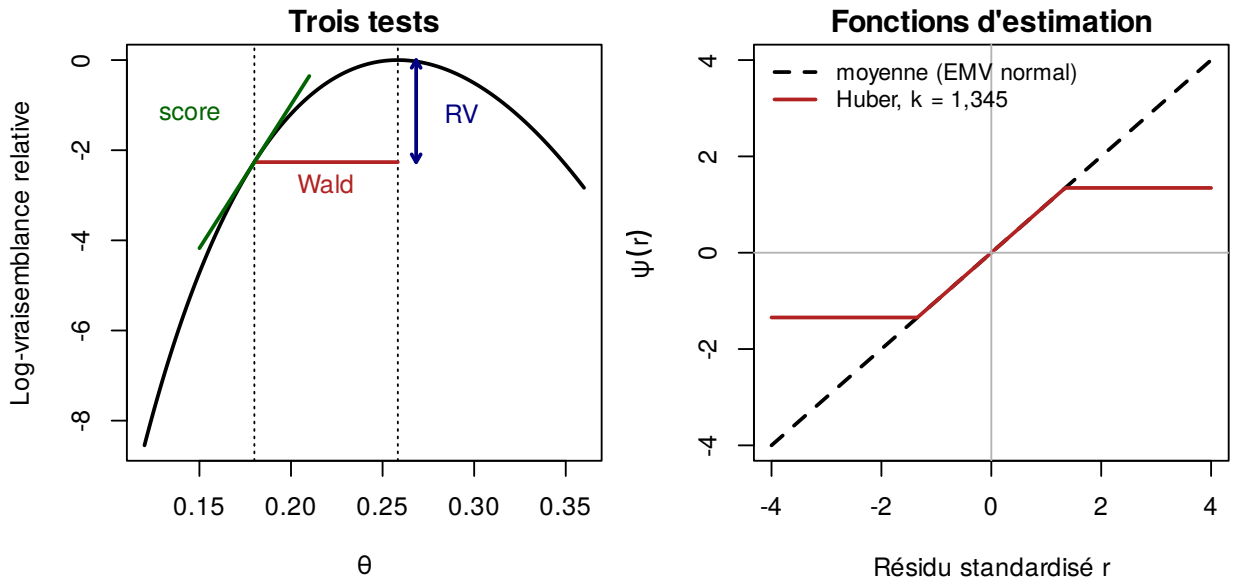


FIGURE 1 – Gauche : géométrie des tests de Wald (écart horizontal), du rapport de vraisemblances (écart vertical) et du score (pente en θ_0) pour l'exemple binomial ; les pointillés verticaux marquent $\theta_0 = 0{,}18$ et $\widehat{\theta} = 0{,}258$. Droite : fonctions d'estimation ψ de la moyenne et de l'estimateur de Huber (section 4).

L'équation est dite *sans biais* pour θ_0 si $\mathrm{E}_F\{\psi(Y, \theta_0)\} = 0$. C'est cette condition qui *définit* la cible : $\theta_0 = \theta(F)$ est la solution de $\mathrm{E}_F\{\psi(Y, \theta)\} = 0$.

Le « M » rappelle que beaucoup de ces estimateurs maximisent (ou minimisent) un critère $\sum_i \rho(Y_i, \theta)$, avec $\psi = \partial\rho/\partial\theta$; mais seule l'équation compte. Exemples :

— *EMV* : $\psi(y, \theta) = \partial \log f_\theta(y)/\partial\theta$; l'équation est sans biais si le modèle est correct (proposition 2.4).

— *Moyenne* : $\psi(y, \mu) = y - \mu$; sans biais pour $\mu = \mathrm{E}_F(Y)$ quelle que soit F.

— *Moyenne et variance* : $\psi(y, \theta) = \big(y - \mu, (y - \mu)^2 - \sigma^2\big)^\top$.

— *Quantile d'ordre* β : $\psi(y, q) = \beta - \mathbf{1}\{y \le q\}$ (fonction non dérivable : la théorie s'étend, avec $A = f_Y(q_\beta)$).

— *Moindres carrés* : $\psi(y, x, \beta) = x(y - x^\top\beta)$; sans biais dès que $\mathrm{E}(Y \mid x) = x^\top\beta$, même si la variance n'est pas constante.

— *Quasi-vraisemblance et GEE* (séance 6) : $\psi_i = D_i^\top V_i^{-1}(Y_i - \mu_i)$, sans biais dès que le modèle de la moyenne est correct, même si la matrice de travail V_i est fausse.

— *Estimateur robuste de Huber* (section 4.5).

4.2 Convergence

Théorème 4.2 (Convergence, version informelle)**.** *Supposons que θ_0 est l'unique solution de $\Psi(\theta) = \mathrm{E}_F\,\psi(Y,\theta) = 0$ et que $\sup_{\theta\in K}\|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{\mathbb{P}} 0$ sur un voisinage compact K de θ_0 (loi des grands nombres uniforme, garantie par exemple si ψ est continue en θ et dominée par une fonction intégrable). Alors il existe une suite de racines $\widehat{\theta}$ de $\Psi_n(\theta) = 0$ telle que $\widehat{\theta} \xrightarrow{\mathbb{P}} \theta_0$.*

L'idée est la même que pour l'EMV : la fonction aléatoire Ψ_n se rapproche uniformément de la fonction déterministe Ψ, dont l'unique zéro est θ_0 ; les zéros de Ψ_n se rapprochent donc de θ_0. Pour l'EMV d'un modèle correct, $\Psi(\theta) = \mathrm{E}_{\theta_0} S_1(\theta)$ s'annule en θ_0 ; pour un modèle mal spécifié, elle s'annule au pseudo-vrai paramètre θ^*.

4.3 Normalité asymptotique et variance sandwich

Théorème 4.3 (Normalité asymptotique d'un estimateur M)**.** *Supposons $\widehat{\theta} \xrightarrow{\mathbb{P}} \theta_0$ avec $\mathrm{E}_F\,\psi(Y,\theta_0) = 0$, ψ deux fois continûment différentiable en θ au voisinage de θ_0 avec dérivées secondes dominées, et définissons*

$$A(\theta_0) = -\,\mathrm{E}_F\Big\{\frac{\partial\psi(Y,\theta)}{\partial\theta^\top}\Big|_{\theta_0}\Big\} \quad (\textit{inversible}), \qquad B(\theta_0) = \mathrm{E}_F\{\psi(Y,\theta_0)\psi(Y,\theta_0)^\top\} \quad (\textit{finie}).$$

Alors

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) = A(\theta_0)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(Y_i,\theta_0) + o_{\mathbb{P}}(1) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0, A^{-1}BA^{-\top}),$$

où $A^{-\top} = (A^{-1})^\top$.

Esquisse. C'est la démonstration du théorème 2.8, sans l'identité de Bartlett. Le développement de Ψ_n autour de θ_0 donne

$$0 = \Psi_n(\widehat{\theta}) = \Psi_n(\theta_0) + \frac{\partial\Psi_n(\theta^*)}{\partial\theta^\top}(\widehat{\theta} - \theta_0),$$

avec θ^* entre θ_0 et $\widehat{\theta}$ (ligne par ligne). Comme $\widehat{\theta} \xrightarrow{\mathbb{P}} \theta_0$, la loi des grands nombres et la condition de domination donnent $-\partial\Psi_n(\theta^*)/\partial\theta^\top \xrightarrow{\mathbb{P}} A(\theta_0)$. Donc

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \{A(\theta_0) + o_{\mathbb{P}}(1)\}^{-1}\sqrt{n}\,\Psi_n(\theta_0).$$

Les termes $\psi(Y_i,\theta_0)$ sont i.i.d., centrés (équation sans biais) et de variance B ; le TCL donne $\sqrt{n}\,\Psi_n(\theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}_p(0,B)$. Par Slutsky, $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} A^{-1}Z$ avec $Z \sim \mathcal{N}(0,B)$, et $\mathrm{Var}(A^{-1}Z) = A^{-1}BA^{-\top}$. □

Estimation de la variance sandwich. On remplace les espérances par des moyennes empiriques évaluées en $\widehat{\theta}$:

$$\widehat{A} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\psi(Y_i,\theta)}{\partial\theta^\top}\Big|_{\widehat{\theta}}, \qquad \widehat{B} = \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i,\widehat{\theta})\psi(Y_i,\widehat{\theta})^\top, \qquad \widehat{\mathrm{Var}}(\widehat{\theta}) = \frac{1}{n}\,\widehat{A}^{-1}\widehat{B}\widehat{A}^{-\top}.$$

On parle de variance « sandwich » ($\widehat{A}^{-1}$ est le pain, $\widehat{B}$ la garniture), ou encore d'estimateur robuste, empirique, de Huber-White. L'estimateur $\widehat{A}^{-1}\widehat{B}\widehat{A}^{-\top}$ est convergent sous des conditions semblables à celles du théorème 4.3 ; à petite taille d'échantillon, il a tendance à sous-estimer la variance (nous le verrons dans l'étude de simulation de la séance 2).

Remarque 4.4. Contrairement à l'information de Fisher, A n'est pas symétrique en général : c'est le cas notamment lorsqu'on empile l'équation d'un paramètre de nuisance et celle du paramètre d'intérêt (estimation en deux étapes, comme aux séances 9 et 10), ou lorsque ψ n'est pas le gradient d'un critère (p. ex. les GEE avec une matrice de corrélation de travail non diagonale et une variance qui dépend de la moyenne). D'où l'écriture $A^{-1}BA^{-\top}$ plutôt que $A^{-1}BA^{-1}$.

4.4 Cas particulier : l'EMV, bien ou mal spécifié

Prenons $\psi = S_1$, le score d'une observation.

— *Modèle correct.* L'identité de Bartlett donne $A = -\operatorname{E}(U_1) = I(\theta_0)$ et $B = \operatorname{E}(S_1 S_1^\top) = I(\theta_0)$, donc $A^{-1}BA^{-\top} = I^{-1}II^{-1} = I^{-1}$: on retrouve le théorème 2.8.

— *Modèle mal spécifié.* $\widehat{\theta}$ converge vers le pseudo-vrai paramètre θ^*, et $A \neq B$ en général. La variance « modèle » $\{nI(\widehat{\theta})\}^{-1}$ est alors fausse, alors que la variance sandwich reste valide pour θ^* (White, 1982). Si θ^* coïncide avec la quantité d'intérêt (p. ex. les coefficients de la moyenne dans un modèle linéaire généralisé dont seule la variance est fausse), on conserve une inférence valide.

Exemple 4.5 (Poisson mal spécifié : calcul explicite)**.** On ajuste un modèle de Poisson de moyenne λ à des comptes Y_i i.i.d. de moyenne μ et de variance σ^2 quelconque (par exemple surdispersés, $\sigma^2 > \mu$). Le score d'une observation est $\psi(y, \lambda) = y/\lambda - 1$, et l'équation est sans biais pour $\lambda = \mu$ quelle que soit la loi : l'EMV $\widehat{\lambda} = \bar{Y}$ converge vers μ. Mais

$$A = -\operatorname{E}\Big(-\frac{Y}{\lambda^2}\Big)\Big|_{\lambda=\mu} = \frac{1}{\mu}, \qquad B = \operatorname{E}\Big(\frac{Y}{\mu} - 1\Big)^2 = \frac{\sigma^2}{\mu^2}, \qquad A^{-1}BA^{-1} = \sigma^2.$$

La variance sandwich σ^2/n est la bonne ($\operatorname{Var}\bar{Y} = \sigma^2/n$), alors que la variance « modèle » $I^{-1}/n = \mu/n$ la sous-estime d'un facteur égal au paramètre de dispersion σ^2/μ.

Exemple 4.6 (Régression de Poisson avec surdispersion)**.** Dans une étude sur les exacerbations d'asthme, Y_i est le nombre d'exacerbations en un an et x_i indique un traitement. On ajuste $\log \operatorname{E}(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$ par le maximum de vraisemblance de Poisson, mais les données sont surdispersées (binomiale négative). Pour ce modèle, $\psi_i = x_i(Y_i - \mu_i)$ (avec $x_i = (1, x_i)^\top$), $\widehat{A} = n^{-1}\sum_i \widehat{\mu}_i x_i x_i^\top$ et $\widehat{B} = n^{-1}\sum_i (Y_i - \widehat{\mu}_i)^2 x_i x_i^\top$.

```
set.seed(15)
n <- 200
x <- rbinom(n, 1, 0.5)
mu <- exp(1 + 0.4 * x)
y <- rnbinom(n, mu = mu, size = 1.5)        # Var(Y) = mu + mu^2/1.5
fit <- glm(y ~ x, family = poisson)
X <- model.matrix(fit); muh <- fitted(fit)
A  <- crossprod(X, X * muh)                 # somme des -d psi / d beta
Bm <- crossprod(X * (y - muh))              # somme des psi psi^T
Vsand <- solve(A) %*% Bm %*% solve(A)
round(cbind(coef = coef(fit), se.modele = sqrt(diag(solve(A))),
            se.sandwich = sqrt(diag(Vsand))), 4)
```

```
             coef se.modele se.sandwich
(Intercept) 0.8607    0.0705      0.1071
x           0.4925    0.0850      0.1391
```

Les erreurs types fondées sur le modèle sont environ 40 % trop petites. Pour vérifier lequel des deux calculs est correct, on répète l'expérience 1000 fois (même plan x, nouvelles réponses) et on compare l'écart type empirique de $\widehat{\beta}_1$ aux moyennes des deux erreurs types (c'est une petite étude de simulation, le sujet de la séance 2) :

```
 sd.empirique   se.modele.moy se.sandwich.moy
       0.1458          0.0811          0.1408
couv.modele couv.sandwich
      0.730         0.949
```

Les intervalles à 95 % fondés sur le modèle ne couvrent la vraie valeur $\beta_1 = 0{,}4$ que 73 % du temps ; ceux fondés sur la variance sandwich ont la couverture nominale. (Les erreurs types sandwich coïncident avec `sandwich::sandwich(fit)` du module `sandwich`, qu'on utilisera en pratique.)

4.5 Exemple : estimation robuste d'une moyenne (Huber)

La moyenne résout $\sum_i (Y_i - \mu) = 0$: chaque observation pèse proportionnellement à son résidu, et une seule valeur aberrante peut déplacer l'estimateur arbitrairement loin. Huber (1964) propose de tronquer la fonction d'estimation :

$$\sum_{i=1}^{n} \psi_k\Big(\frac{Y_i - \mu}{s}\Big) = 0, \qquad \psi_k(r) = \max\{-k, \min(k, r)\} = \begin{cases} r & |r| \le k, \\ k\,\mathrm{signe}(r) & |r| > k, \end{cases}$$

où s est une échelle robuste fixée au préalable (souvent l'écart absolu médian normalisé, `mad` en R) et k une constante de réglage (figure 1, droite). Si la loi de Y est symétrique autour de μ_0, l'équation est sans biais pour μ_0 (la moyenne et la médiane). Les quantités du théorème 4.3 sont

$$A = \frac{1}{s}\,\mathbb{P}(|Y - \mu_0| \le ks), \qquad B = \mathrm{E}\Big\{\psi_k\Big(\frac{Y - \mu_0}{s}\Big)^2\Big\}, \qquad \mathrm{Var}(\widehat{\mu}) \approx \frac{B}{nA^2}.$$

(La fonction ψ_k n'est pas dérivable en $\pm k$, mais elle l'est presque partout et la théorie s'applique.) Sous la loi normale avec $s = \sigma$, l'efficacité relative à la moyenne vaut $\mathbb{P}(|Z| \le k)^2 / \mathrm{E}\{\psi_k(Z)^2\}$, soit 0,95 pour $k = 1{,}345$: on paie 5 % de variance quand les données sont normales, en échange d'une grande stabilité quand elles ne le sont pas. L'estimateur se calcule par moindres carrés repondérés itératifs : avec $w_i = \psi_k(r_i)/r_i = \min(1, k/|r_i|)$, l'équation s'écrit $\sum_i w_i(Y_i - \mu) = 0$, d'où $\mu \leftarrow \sum_i w_i Y_i / \sum_i w_i$.

Exemple 4.7. Cinquante mesures de cholestérol HDL (mmol/L), dont trois erreurs de saisie (virgule décimale déplacée).

```r
huber_M <- function(y, k = 1.345, s = mad(y), tol = 1e-10) {
  mu <- median(y)
  repeat {
    r <- (y - mu) / s
    w <- pmin(1, k / abs(r))              # poids IRLS
    mu.new <- sum(w * y) / sum(w)
    if (abs(mu.new - mu) < tol) break
    mu <- mu.new
  }
  r <- (y - mu) / s
  A <- mean(abs(r) <= k) / s              # estimation de A
  B <- mean(pmax(-k, pmin(k, r))^2)       # estimation de B
  c(mu = mu, se = sqrt(B / A^2 / length(y)))
```

```
}
set.seed(2027)
hdl <- c(round(rnorm(47, 1.4, 0.3), 2), 14.2, 13.8, 12.9)
round(c(moyenne = mean(hdl), se.moy = sd(hdl)/sqrt(50), mediane = median(hdl)), 3)
round(huber_M(hdl), 3)
MASS::huber(hdl, k = 1.345)$mu             # vérification
```

```
moyenne  se.moy mediane
  2.187   0.415   1.485
   mu    se
1.485 0.049
[1] 1.485432
```

Trois valeurs aberrantes font passer la moyenne de 1,4 à 2,19 et multiplient son erreur type par près de dix ; l'estimateur de Huber reste près du centre des données valides et son erreur type sandwich (0,049) est du même ordre que $0{,}3/\sqrt{47} \approx 0{,}044$. Notre erreur type ignore la variabilité de s : c'est sans conséquence asymptotique lorsque la loi est symétrique (l'équation conjointe en (μ, s) a alors une matrice A bloc-diagonale), mais il faut empiler les deux équations sinon.

4.6 Fonction d'influence et estimateurs asymptotiquement linéaires

Le théorème 4.3 contient plus que la loi limite : il affirme que l'estimateur se comporte, au premier ordre, comme une *moyenne* de termes i.i.d.

Définition 4.8. Un estimateur $\widehat{\theta}$ de $\theta_0 = \theta(F)$ est *asymptotiquement linéaire* de fonction d'influence $\mathrm{IF}(y)$ si

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathrm{IF}(Y_i) + o_{\mathbb{P}}(1), \qquad \mathrm{E}_F\{\mathrm{IF}(Y)\} = 0, \quad \mathrm{E}_F\{\mathrm{IF}(Y)\,\mathrm{IF}(Y)^\top\} < \infty.$$

Le TCL donne alors immédiatement $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{E}\{\mathrm{IF}(Y)\,\mathrm{IF}(Y)^\top\})$.

Pour un estimateur M, le théorème 4.3 donne

$$\boxed{\mathrm{IF}(y) = A(\theta_0)^{-1}\psi(y, \theta_0)} \qquad \text{et} \qquad \mathrm{E}(\mathrm{IF}\ \mathrm{IF}^\top) = A^{-1}BA^{-\top}.$$

Exemples : pour la moyenne, $\mathrm{IF}(y) = y - \mu$; pour l'EMV d'un modèle correct, $\mathrm{IF}(y) = I(\theta_0)^{-1} S(y; \theta_0)$; pour l'estimateur de Huber, $\mathrm{IF}(y) = s\,\psi_k\{(y - \mu_0)/s\}/\mathbb{P}(|Y - \mu_0| \le ks)$, une fonction *bornée*.

Trois usages de la fonction d'influence reviendront dans le cours :

1. *Variance.* La variance asymptotique est $\mathrm{E}(\mathrm{IF}\ \mathrm{IF}^\top)$, estimée par $n^{-1}\sum_i \widehat{\mathrm{IF}}(Y_i)\widehat{\mathrm{IF}}(Y_i)^\top$: c'est ainsi qu'on obtient la variance des GEE (séance 6), des estimateurs pondérés par l'inverse de la probabilité (séance 9) ou de l'estimation ciblée (séance 7).
2. *Composition.* Si $\widehat{\theta}$ a pour fonction d'influence IF_θ, alors $g(\widehat{\theta})$ a pour fonction d'influence $\nabla g(\theta_0)\ \mathrm{IF}_\theta$: c'est la méthode delta réécrite en termes de linéarisation. De même, un estimateur qui dépend d'un paramètre de nuisance estimé voit sa fonction d'influence corrigée d'un terme dû à cette estimation (cas de l'échantillonnage à deux phases, séance 10).

3. *Robustesse.* La fonction d'influence mesure l'effet d'une observation y sur l'estimateur : une IF non bornée (moyenne, EMV normal) signifie qu'une seule valeur aberrante peut tout faire basculer ; borner ψ borne l'influence. La question d'examen « donnez deux façons de modifier la fonction d'influence $y - \mu$ de la moyenne pour obtenir une estimation robuste » (examen final 2023) s'y rattache directement.

Niveau doctorat : la fonction d'influence comme dérivée de Gâteaux

Hampel (1974) définit la fonction d'influence d'une fonctionnelle T (avec $\theta_0 = T(F)$) par

$$\mathrm{IF}(y; T, F) = \lim_{\epsilon \downarrow 0} \frac{T\{(1-\epsilon)F + \epsilon\delta_y\} - T(F)}{\epsilon},$$

où δ_y est la masse ponctuelle en y : c'est l'effet relatif d'une contamination infinitésimale en y. Pour un estimateur M, $T(F_\epsilon)$ résout $\int \psi\{u, T(F_\epsilon)\}\,\mathrm{d}F_\epsilon(u) = 0$, soit $(1-\epsilon)\,\mathrm{E}_F\,\psi\{Y, T(F_\epsilon)\} + \epsilon\,\psi\{y, T(F_\epsilon)\} = 0$. En dérivant en $\epsilon = 0$ et en utilisant $\mathrm{E}_F\,\psi(Y, \theta_0) = 0$: $-A\ \mathrm{IF}(y) + \psi(y, \theta_0) = 0$, donc $\mathrm{IF}(y) = A^{-1}\psi(y, \theta_0)$, ce qui coïncide avec le terme de linéarisation. Un développement de von Mises, $T(\widehat{F}_n) - T(F) = \int \mathrm{IF}\ \mathrm{d}(\widehat{F}_n - F) + \text{reste} = n^{-1}\sum_i \mathrm{IF}(Y_i) + \text{reste}$, relie les deux points de vue. Dans la théorie semi-paramétrique (Tsiatis, 2006, ch. 3 ; séance 6), les estimateurs réguliers et asymptotiquement linéaires sont caractérisés par leurs fonctions d'influence, et l'estimateur efficace est celui dont la fonction d'influence est de variance minimale.

4.7 Lien avec le manuel

Wu (2010, § 9.2.2) présente les estimateurs M dans le contexte de l'estimation robuste des modèles à effets mixtes : une fonction ρ (ou ψ) bornée remplace la contribution quadratique de chaque résidu, et la variance des estimateurs prend la forme sandwich vue ici. Nous utiliserons cette lecture à la séance 6, avec l'exemple `sleepstudy` ajusté de façon robuste. Lisez cette sous-section en identifiant, dans les notations du manuel, la fonction ψ, les matrices A et B, et la fonction d'influence.

Grille méthodologique : estimateurs M

1. **Objet d'intérêt** : $\theta_0 = \theta(F)$, défini par $\mathrm{E}_F\,\psi(Y, \theta_0) = 0$ (paramètre d'un modèle, pseudo-vrai paramètre, moyenne, quantile, coefficients d'une moyenne conditionnelle…).
2. **Nuisance et biais** : la loi F elle-même n'est pas modélisée en entier ; biais si l'équation n'est pas centrée en la quantité d'intérêt (p. ex. modèle de la moyenne faux, loi asymétrique pour Huber).
3. **Approche** : choisir ψ pour que l'équation reste sans biais sous des hypothèses faibles (seulement la moyenne, symétrie…) et éventuellement pour borner l'influence.
4. **Critère** : résoudre $\sum_i \psi(Y_i, \theta) = 0$.
5. **Calcul** : Newton-Raphson sur Ψ_n, moindres carrés repondérés itératifs.
6. **Inférence** : $\widehat{\theta} \approx \mathcal{N}\big(\theta_0, n^{-1}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-\top}\big)$; tests de Wald « robustes » fondés sur cette variance.

À retenir

- Un estimateur M résout $\sum_i \psi(Y_i, \theta) = 0$; l'équation sans biais $\mathrm{E}_F\, \psi(Y, \theta_0) = 0$ définit la cible.
- $\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, A^{-1}BA^{-\top})$ avec $A = -\mathrm{E}\, \partial\psi/\partial\theta^\top$ et $B = \mathrm{E}\, \psi\psi^\top$; on estime A et B par des moyennes empiriques.
- EMV d'un modèle correct : $A = B = I$. Modèle mal spécifié : $A \neq B$, seule la variance sandwich est valide.
- Linéarisation : $\widehat{\theta} - \theta_0 \approx n^{-1} \sum_i \mathrm{IF}(Y_i)$ avec $\mathrm{IF} = A^{-1}\psi$; variance $= \mathrm{E}(\mathrm{IF}\ \mathrm{IF}^\top)$; ψ bornée $\Rightarrow$ estimateur robuste.

5 Exercices

Les exercices 5.1 à 5.3 sont repris de la série « Rappels – Exercices » (H26, L. Lakhal-Chaieb). Les exercices 5.4 et 5.5 sont nouveaux (H27) et portent sur les estimateurs M.

Exercice 5.1 (Rappels H26, exercice 1)**.** Soit $X_1, \ldots, X_n$ un échantillon de la loi uniforme continue sur $(-1, 1)$. On pose $Z_1 = n^{-1} \sum_i X_i$, $Z_2 = n^{-1} \sum_i X_i^2$ et $Z = \sum_i X_i / \sum_i X_i^2$.

a) Trouver la loi asymptotique conjointe de $\sqrt{n}\,(Z_1, Z_2 - 1/3)^\top$.

b) Trouver la loi asymptotique de $\sqrt{n}\, Z$.

Exercice 5.2 (Rappels H26, exercice 2)**.** Soit $X_1, \ldots, X_n$ un échantillon de la loi $\mathcal{N}(\mu, \sigma^2)$. Trouver la loi asymptotique de $\sqrt{n}\,(S_n^2 - \sigma^2)$, où S_n^2 est la variance échantillonnale. Que devient le résultat si la loi n'est pas normale mais possède un quatrième moment fini ?

Exercice 5.3 (Rappels H26, exercice 3)**.** Soit $X_1, \ldots, X_n$ un échantillon de densité $f_\theta(x) = \theta x^{\theta-1} \mathbf{1}\{x \in [0, 1]\}$, $\theta > 0$.

a) Montrer que $\mathrm{E}(X) = \theta/(\theta + 1)$.

b) En déduire $\widetilde{\theta}$, l'estimateur des moments de θ.

c) Donner la loi asymptotique de $\sqrt{n}(\widetilde{\theta} - \theta)$.

d) Trouver $\widehat{\theta}$, l'estimateur du maximum de vraisemblance de θ.

e) Donner la loi asymptotique de $\sqrt{n}(\widehat{\theta} - \theta)$.

f) Lequel des deux estimateurs possède la plus petite variance asymptotique ?

Exercice 5.4 (Nouveau, H27)**.** Montrer que la médiane empirique est (approximativement) un estimateur M avec $\psi(y, q) = \frac{1}{2} - \mathbf{1}\{y \leq q\}$. En admettant que le théorème 4.3 s'applique avec $A = f_Y(q_{1/2})$, donner sa variance asymptotique, puis l'efficacité de la médiane relativement à la moyenne lorsque $Y \sim \mathcal{N}(\mu, \sigma^2)$. Donner sa fonction d'influence et commenter.

Exercice 5.5 (Nouveau, H27)**.** On ajuste un modèle exponentiel de taux λ à des temps de survie $Y_1, \ldots, Y_n$ i.i.d. qui suivent en réalité une loi de Weibull de forme $\alpha \neq 1$ et d'échelle 1.

a) Vers quoi converge l'EMV $\widehat{\lambda} = 1/\bar{Y}$? Est-ce une quantité interprétable ?

b) Calculer A et B pour la fonction $\psi(y, \lambda) = 1/\lambda - y$, et comparer la variance sandwich à la variance « modèle » λ^2/n. Pour quelles valeurs de α la variance modèle est-elle trop petite ?

c) (Lien avec la séance 2) Proposer une petite étude de simulation qui vérifie votre réponse en b).

Références

— Boos, D. D. et Stefanski, L. A. (2013). *Essential Statistical Inference : Theory and Methods*. Springer. (Ch. 7 : estimateurs M.)

— Casella, G. et Berger, R. L. (2002). *Statistical Inference*, 2e éd. Duxbury. (Ch. 5 et 10 : convergence, méthode delta, propriétés asymptotiques, tests.)

— Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383–393.

— Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35, 73–101.

— Stefanski, L. A. et Boos, D. D. (2002). The calculus of M-estimation. *The American Statistician*, 56, 29–38.

— Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer.

— van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. (Ch. 2–5.)

— White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.

— Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman & Hall/CRC. (§ 9.2.2.)

— Lakhal-Chaieb, L. (2026). *STT-7360 : Rappels* et *Rappels – Exercices – Solutions*. Notes de cours, Université Laval.