

Statistiques et Analyse de Données

Paul Lehaut

October 30, 2025

Contents

1 Rappels	3
1.1 Espérance et (Co)Variance	3
1.2 Indépendance	3
1.3 Variables Aléatoires Discrètes	4
1.4 Variables Aléatoires à Densité	4
1.5 Somme de VAs indépendantes	4
1.6 Fonction de Répartition	4
1.7 Fonction Charactéristique	5
1.8 Vecteurs Gaussiens	5
1.9 Théorèmes de Convergence	6
2 Estimateur dans un Modèle Paramétrique	7
2.1 Estimateurs	7
2.2 Biais et Risque Quadratique	7
2.3 Modèle Paramétrique et Estimation du Moment	8
2.4 Convergence Normale	9
3 Statistiques dans des Modèles Gaussiens	9
3.1 Statistiques d'Echantillons Gaussiens	10
3.2 La Distribution de Student	10
3.3 Régression Linéaire avec Erreurs Gaussiennes	10
4 Intervalles de Confiance	11
4.1 Définitions Générales	11
4.2 Construction d'Intervalles de Confiance Exact	11
4.2.1 Fonction Pivot	11
4.2.2 Exemple dans le Modèle Gaussien	11
4.2.3 Résumé de la Méthode	12
4.3 Construction d'Intervalle de Confiance Asymptotique	12
4.4 Construction d'Intervalle de Confiance par Excès	13
4.4.1 L'Inégalité de Bienaymé-Chebychev	13
4.4.2 Inégalité de Hoeffding	13
5 Estimateur du Maximum de Vraisemblance	13
5.1 Vraisemblance d'un Echantillon et Estimateur	14
5.2 Exemples	14
5.2.1 Modèle de Bernoulli	14
5.2.2 Cas du Modèle Gaussien	15
5.3 Optimalité du MLE	15
5.3.1 Modèle Régulier et Information de Fisher	15
5.3.2 Estimateur Efficace	15

1 Rappels

1.1 Espérance et (Co)Variance

On considère dans cette section des VAR.

Définition:

L'espérance de $X \sim \mathbb{P}$ est l'intégral de Lebesgue:

$$\mathbb{E}(X) := \int_{\omega \in \Omega} X(\omega) d\mathbb{P}(\omega) = \int_{x \in \mathbb{R}} x d\mathbb{P}(x)$$

qui est bien définie si X est intégrable, c'est-à-dire si $\mathbb{E}(|X|) < +\infty$.

Si par ailleurs $\mathbb{E}(X^2) < +\infty$, alors la variance de X est bien définie par:

$$\mathbb{V}(X) := \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2.$$

L'inégalité de Jensen donne alors, pour X une VAR d intégrable, soit f une fonction à valeurs réelles convexe définie sur \mathbb{R}^d , alors $\mathbb{E}(f(X))$ est bien définie et:

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

Définition:

La covariance entre X et Y , telles que $\mathbb{E}(X) < +\infty$ et $\mathbb{E}(Y) < +\infty$, est définie par:

$$\text{Cov}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

on a alors l'identité remarquable suivante:

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + 2\text{Cov}(X, Y) + \mathbb{V}(Y).$$

Lorsque X est une VAR d de carré intégrable, on peut définir sa matrice de covariance qui est symétrique positive.

Définition: Le coefficient de corrélation de X et Y est définie par:

$$\rho_{X,Y} := \begin{cases} \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}} \in [-1, 1] & \text{si } \text{Var}(X) \text{Var}(Y) > 0, \\ 0 & \text{sinon.} \end{cases}$$

1.2 Indépendance

Définition: Une famille finie de variables aléatoires (X_1, \dots, X_n) définies sur (Ω, \mathcal{F}) est indépendante si pour toute famille de sous-ensembles mesurables (C_1, \dots, C_n) on a:

$$\mathbb{P}(X_1 \in C_1, \dots, X_n \in C_n) = \mathbb{P}(X_1 \in C_1) \dots \mathbb{P}(X_n \in C_n)$$

ou, de façon équivalente, si:

$$\mathbb{P} = \mathbb{P}_{X_1} \otimes \dots \otimes \mathbb{P}_{X_n}.$$

Si X_1, \dots, X_n sont indépendantes alors, pour toutes fonctions mesurables f_1, \dots, f_n telles que $f_1(X_1), \dots, f_n(X_n)$ soient intégrables, on a:

$$\mathbb{E}(f_1(X_1), \dots, f_n(X_n)) = \mathbb{E}(f_1(X_1)) \dots \mathbb{E}(f_n(X_n)).$$

1.3 Variables Aléatoires Discrètes

Si E est discret, toutes mesures de probabilités est caractérisée par la famille de nombres $(p(x), x \in E)$. Les variables aléatoires principales sont définies dans le polycopié.

1.4 Variables Aléatoires à Densité

Les intégrales sont ici à considérer au sens de Lebesgue.

Définition:

Une variable aléatoire X est dite à densité p si:

$$\forall C \in \mathcal{B}(\mathbb{R}^d), \mathbb{P}(X \in C) = \int_{x \in C} p(x)dx$$

une fonction mesurable et positive p est une densité de probabilité si et seulement si:

$$\int_{x \in \mathbb{R}^d} p(x)dx = 1.$$

Des variables aléatoires X_1, \dots, X_n sont indépendantes si le vecteur aléatoire (X_1, \dots, X_n) a pour densité $p = p_{X_1} \otimes \dots \otimes p_{X_n}$.

Théorème: Formule de transfert

Si X a une densité p , alors pour toutes fonctions mesurables f telle que $\mathbb{E}(|f(X)|) < +\infty$, alors:

$$\mathbb{E}(f(X)) = \int_{x \in \mathbb{R}^d} f(x)p(x)dx.$$

Les exemples principaux de variables aléatoires à densité se trouvent dans le polycopié.

1.5 Somme de VAs indépendantes

Soient X et Y deux VAs indépendantes, on note $Z := X + Y$.

Proposition:

La densité de Z est:

$$r(z) = \int_{x \in \mathbb{R}^d} p_X(x)p_Y(z - x)dx$$

on l'appelle la convolution de p_X et p_Y .

1.6 Fonction de Répartition

On se place dans le cas où $E = \mathbb{R}$.

Définition:

La fonction de répartition d'un variable aléatoire X est définie par:

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}(X \leq x)$$

alors, pour tout réel $r \in (0, 1)$, un quantile d'ordre r pour X est un nombre q_r tel que:

$$\mathbb{P}(X \leq q_r) = F(q_r) = r.$$

En général un quantile n'existe pas toujours ou n'est pas unique, néanmoins, lorsque X possède une densité qui est positive alors le quantile existe et est unique.

Proposition:

Si X est une VAR à densité p , alors sa fonction de répartition est continue et dérivable presque partout, sa dérivée est presque partout égale à sa fonction de densité.

1.7 Fonction Charactéristique

Soient X et Y des vecteurs aléatoires.

Définition: Fonction Charactéristique

La fonction caractéristique de X est la fonction $\Psi_X : \mathbb{R}^d \rightarrow \mathbb{C}$ définie par:

$$\Psi_X(u) := \mathbb{E}(e^{i\langle u, X \rangle}) = \mathbb{E}(\cos(\langle u, X \rangle)) + i\mathbb{E}(\sin(\langle u, X \rangle)).$$

Proposition:

Si, pour tout $u \in \mathbb{R}^d$, $\Phi_X(u) = \Phi_Y(u)$, alors X et Y sont de même loi.

Proposition:

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ et $Y \sim \mathcal{N}(\nu, \tau^2)$ sont indépendantes, alors:

$$X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2).$$

1.8 Vecteurs Gaussiens

Soit X un vecteur aléatoire.

Définition: Vecteur Gaussien

X est un vecteur Gaussien si, pour tout vecteur u , la variable aléatoire $\langle u, X \rangle$ est gaussienne.

De ce qui précède, on déduit que, en notant $\mathbb{E}(X) = m$ et $Cov(X) = K$, alors:

$$\langle u, X \rangle \sim \mathcal{N}(\langle u, m \rangle, \langle u, Ku \rangle) \text{ et } \Phi_X(u) = \exp(i\langle u, m \rangle - \frac{1}{2}\langle u, Ku \rangle).$$

Par ailleurs, si K est inversible, alors X a la densité:

$$\frac{1}{\sqrt{(2\pi)^d \det K}} \exp\left(-\frac{1}{2} \langle x - m, K^{-1}(x - m) \rangle\right)$$

sinon X n'a pas de densité.

1.9 Théorèmes de Convergence

Soient (X_n) et X des vecteurs aléatoires.

Définition:

(X_n) converge vers X presque sûrement si: $\mathbb{P}(\lim X_n = X) = 1$.

(X_n) converge vers X en probabilité si, pour tout $\epsilon > 0$, $\lim \mathbb{P}(\|X_n - X\| \geq \epsilon) = 0$.

(X_n) converge vers X en distribution si, pour toute fonction continue et bornée $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $\mathbb{E}(f(X_n))$ converge vers $\mathbb{E}(f(X))$.

Théorème: Convergence Dominée

Supposons que (X_n) converge vers X presque sûrement et qu'il existe Y positive et intégrable telle que: $\|X_n\| \leq Y$ presque sûrement, alors $\mathbb{E}(X_n)$ converge vers $\mathbb{E}(X)$.

Pour f une fonction continue, si (X_n) converge presque sûrement (respectivement en probabilité ou en distribution) vers X alors $f(X_n)$ converge presque sûrement vers $f(X)$ (respectivement en probabilité ou en distribution).

Proposition:

(X_n) converge en distribution (c'est-à-dire en loi) vers X si et seulement si $\Psi_{X_n}(u)$ converge vers $\Psi_X(u)$ pour tout u .

La convergence presque sûre implique la converge en probabilité qui implique elle même la converge en distribution.

Si on se place dans le cadre réel, alors on a les équivalences suivantes:

- (X_n) converge vers X en distribution

- $\mathbb{P}(X_n \leq x) \rightarrow \mathbb{P}(X \leq x)$ pour tout x tel que $\mathbb{P}(X = x) = 0$

- $\mathbb{P}(X_n < x) \rightarrow \mathbb{P}(X < x)$ pour tout x tel que $\mathbb{P}(X = x) = 0$

On dit que la suite (X_n) est indépendante et indentiquement distribuée si ses variables sont indépendantes et de même loi. On note alors:

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

la moyenne empirique de X_1, \dots, X_n .

Théorème: La Loi Forte des Grands Nombres

Soit (X_n) une suite de VAR^d iid telle que $\mathbb{E}(\|X_1\|) < +\infty$, alors:

$$\lim \bar{X}_n = \mathbb{E}(X_1) \text{ presque sûrement.}$$

Théorème: Théorème Central Limite Multivarié

Soit (X_n) une suite de VAR^d iid telle que $\mathbb{E}(\|X_1\|^2) < +\infty$, alors:

$$\lim \sqrt{n}(\bar{X}_n - \mathbb{E}(X_1)) = \mathcal{N}(0, \text{Cov}(X_1)) \text{ en distribution.}$$

2 Estimateur dans un Modèle Paramétrique

On considère un échantillon X_1, \dots, X_n de VA iid dans un espace mesurable (E, \mathcal{E}) de loi \mathbb{P} inconnue qu'on cherche à éclaircir. On cherche par exemple à estimer $\mathbb{E}_{\mathbb{P}}(X_1)$, $V_{\mathbb{P}}(X_1)$ l'histogramme de \mathbb{P} ou encore d'autre quantité d'intérêt (QI).

On se restreint à des lois d'une certaine forme caractérisées par un ou certains paramètres. Formellement, on considère une famille de lois: $\{\mathbb{P}_{\theta}; \theta \in \Theta\}$.

2.1 Estimateurs

On cherche donc à estimer une QI à l'aide de notre échantillon de VA, pour ce faire on va chercher à approcher QI par une fonction de l'échantillon appelée statistique.

Définition:

Une statistique T_n est une VA de la forme $T_n = t_n(X_1, \dots, X_n)$ avec t_n déterministe et qui ne dépend pas de \mathbb{P} .

On appelle un estimateur une statistique qui vise à approcher une certaine QI.

Définition:

Un estimateur Z_n d'une QI est dit constant si Z_n converge en probabilité vers QI, il est fortement constant s'il converge vers QI presque sûrement.

Par exemple, dans le cas où $E = \mathbb{R}$, alors la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur fortement constant de $\mathbb{E}(X_1)$.

2.2 Biais et Risque Quadratique

Le biais et la MSE permettent de quantifier la distance d'un estimateur à sa QI dans un régime non asymptotique (ie n est fini).

Définition:

Soit un estimateur intégrable Z_n , le biais de Z_n est:

$$b(Z_n) = \mathbb{E}(Z_n) - QI, \text{ il s'agit de la distance moyenne de l'estimateur à la QI}$$

si ce biais est nul, on dit que l'estimateur est non biaisé.

Il est intéressant de remarquer que la variance empirique est biaisée ce qui motive la définition suivante:

Définition:

Si $E = \mathbb{R}$, l'estimateur non-biaisé de la variance est:

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

Une mesure plus précise de la distance de l'estimateur à la QI peut être donnée par la MSE.

Définition:

Soit Z_n un estimateur de carré intégrable, la MSE de Z_n est définie par:

$$MSE(Z_n) = E(\|Z_n - QI\|^2).$$

Proposition:

On peut a l'égalité suivante:

$$MSE(Z_n) = ||b(Z_n)||^2 + V(Z_n).$$

En général on ne peut pas minimiser à la fois la variance et le biais. En data science il peut être intéressant d'introduire un biais pour réduire la variance du modèle et donc le risque d'overfitting.

2.3 Modèle Paramétrique et Estimation du Moment

On s'intéresse ici à l'estimation de la distribution complète \mathbb{P} de X_1 . Il y a deux approches principales: la méthode non paramétrique (par histogramme) et la méthode paramétrique qui repose sur la supposition que \mathbb{P} a une certaine forme (comme exponentielle ou gaussienne).

Définition:

Un modèle paramétrique sur E est un ensemble de mesures de probabilités:

$$\mathcal{P} = \{\mathbb{P}_\theta; \theta \in \Theta\}$$

sur l'espace E , indexé par un ensemble de paramètres $\Theta \subset \mathbb{R}^k$.

Il est important de constater que si, pour deux valeurs distinctes θ et θ' , on a $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ alors il n'est pas possible de distinguer θ de θ' par simple observation de X_1, \dots, X_n . On travaillera donc toujours en supposant que la fonction $\theta \mapsto \mathbb{P}_\theta$ est injective, on dira alors que \mathcal{P} est identifiable.

On fixe désormais un modèle paramétrique \mathcal{P} . Pour tout θ il est pratique de dénoté par \mathbb{P}_θ la mesure de probabilité pour laquelle, pour tout $n \geq 1$, les VAs X_1, \dots, X_n sont iid selon \mathbb{P}_θ . On définit de façon similaire \mathbb{E}_θ , V_θ etc.

Proposition:

Soit X une VA \mathbb{R} non déterministe intégrable et qui prend ses valeurs dans un intervalle I . Soit $\Phi : I \rightarrow \mathbb{R}$ strictement convexe, alors: $\Phi(\mathbb{E}(X)) < \mathbb{E}(\Phi(X))$.

La méthode des moments est une procédure naturelle pour construire des estimateurs, on détaille ici l'exemple pour le modèle exponentielle: $\{\epsilon(\lambda); \lambda > 0\}$ dans lequel on cherche un estimateur de λ :

- La loi forte des grands nombres donne, pour tout $\lambda > 0$: $\bar{X}_n \rightarrow \mathbb{E}_\lambda(X_1) = \frac{1}{\lambda}$ presque sûrement, donc $\frac{1}{\bar{X}_n} \rightarrow \lambda$ presque sûrement.
La continuité de la fonction $x \mapsto \frac{1}{x}$ sur \mathbb{R}_+^* assure donc que $\bar{\lambda}_n = \frac{1}{\bar{X}_n}$ est un estimateur fortement consistant de λ .

La généralisation abstraite de cette méthode s'énonce de la façon suivante:

- Pour l'estimation de $g(\theta) \in \mathbb{R}^d$, cela consiste à trouver ϕ et m des fonctions telles que:

$$\forall \theta \in \Theta, \mathbb{E}_\theta(\phi(X_1)) = m(g(\theta)).$$

Pour le modèle exponentielle on a pris: $\phi(x) = x$, $g(\lambda) = \lambda$ et $m(\lambda) = \frac{1}{\lambda}$.

Alors, la loi forte des grands nombres nous permet donc d'approximer $m(g(\theta))$ par: $\frac{1}{n} \sum_{i=1}^n \phi(X_i)$ de sorte que, si m possède une fonction réciproque continue m^{-1} , alors:

$$Z_n = m^{-1}\left(\frac{1}{n} \sum_{i=1}^n \phi(X_i)\right)$$

est un estimateur fortement consistant de $g(\theta)$.

2.4 Convergence Normale

La construction d'un estimateur par la méthode des moments dépend du choix arbitraire de la fonction ϕ , donc différents choix de ϕ peuvent donner différents estimateurs qui sont tous, par construction, fortement consistant. Pour déterminer l'estimateur le plus intéressant en pratique, on peut s'intéresser à celui qui converge 'le plus vite' vers la QI. Cette vitesse de convergence peut être mesurée à l'aide de la notion de variance asymptotique.

Définition:

Un estimateur consistant Z_n de $g(\theta)$ est asymptotiquement normal si, pour tout θ , il existe une matrice symétrique positive $K(\theta) \in \mathcal{M}_d(\mathbb{R})$ telle que: $\sqrt{n}(Z_n - g(\theta))$ converge en distribution vers $\mathcal{N}_d(0, K(\theta))$.

La fonction $\theta \mapsto K(\theta)$ est appelé la covariance asymptotique de Z_n .

Théorème: Méthode delta

Soit (ζ_n) une suite de VA à valeurs dans \mathbb{R}^k et $a \in \mathbb{R}^k$ telles que $\zeta_n \rightarrow a$ en probabilité et $\sqrt{n}(\zeta_n - a)$ converge en distribution vers un vecteur aléatoire $Y \in \mathbb{R}^k$. Soit \mathcal{U} un ouvert de \mathbb{R}^k qui contient a , soit $\Phi : \mathcal{U} \rightarrow \mathbb{R}^d$ de classe \mathcal{C}^1 , alors:

$$\lim_{n \rightarrow +\infty} \sqrt{n}(\Phi(\zeta_n) - \Phi(a)) = \nabla \Phi(a)Y$$

en distribution.

Théorème: Slutsky

Soit $((X_n, Y_n))$ une suite de couples de VAs telle que (X_n) converge en probabilité vers une variable déterministe a et que (Y_n) converge en distribution vers une variable aléatoire Y . Alors $((X_n, Y_n))$ converge en distribution vers (a, Y) , et, par conséquent, pour toute fonction continue Ψ , $(\Psi(X_n, Y_n))$ converge en distribution vers $\Psi(a, Y)$.

3 Statistiques dans des Modèles Gaussiens

On commence par quelques rappels.

Définition:

Un vecteur aléatoire $G \in \mathbb{R}^n$ est un vecteur gaussien standard si chacune de ces composantes est de loi $\mathcal{N}(0, 1)$ et si elles sont indépendantes.

La fonction caractéristique d'un vecteur gaussien standard est: $\psi_G(u) = e^{-\frac{\|u\|^2}{2}}$ (il s'agit du cas particulier de la fonction caractéristique d'un vecteur gaussien: $\phi_X(x) = \exp(it^T m - \frac{1}{2}t^T \Sigma t)$).

Théorème: Cochran

Soit $G \sim \mathcal{N}_n(0, I_n)$, pour tout sous-espace-vectoriel E de \mathbb{R}^n , les coordonnées de G dans toutes base orthonormée de E forment un vecteur gaussien standard.

Définition:

Pour $n \geq 1$, la distribution χ -carré avec n degrés de liberté, notée $\chi_2(n)$, est la loi de la VA:

$$Z_n = \sum_{i=1}^n G_i^2 = \|G\|^2 \text{ avec } G \sim \mathcal{N}_n(0, I_n).$$

Son espérance est n , sa variance $2n$.

3.1 Statistiques d'Echantillons Gaussiens

On note $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_i)^2$ l'estimateur non biaisé de la variance.

Proposition:

En considérant $\mathbb{P}_{\mu, \sigma^2}$, alors les estimateurs \bar{X}_n et S_n^2 sont indépendants et:

$$\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \text{ et } (n-1) \frac{S_n^2}{\sigma^2} \sim \chi_2(n-1).$$

On introduit pour la suite la variable aléatoire réduite $X'_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$ et on définit comme on s'y attend \bar{X}'_n et S'^2_n telles que:

$$\bar{X}_n = \mu + \sigma \bar{X}'_n \text{ et } S_n^2 = \sigma^2 S'^2_n$$

alors, en notant $E_1 = \text{Vect}\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right)$, $E_2 = E_1^\perp$ et $e = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, il vient:

$$G_{E_1} = \langle G, e \rangle e = \bar{X}'_n \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ et } \|G_{E_2}\|^2 = (n-1)(S'_n)^2.$$

3.2 La Distribution de Student

Soit $n \geq 1$, alors:

Définition:

La distribution de Student avec n degrés de liberté, notée $t(n)$, est la loi de la VA: $T_n = Y \sqrt{\frac{n}{Z_n}}$ avec $Z_n \sim \chi_2(n)$ indépendante de $Y \sim \mathcal{N}(0, 1)$.

Proposition:

Pour $\mathbb{P}_{\mu, \sigma^2}$, alors: $\frac{\bar{X}_n - \mu}{\sqrt{\frac{S_n^2}{n}}} \sim t(n-1)$.

3.3 Régression Linéaire avec Erreurs Gaussiennes

On s'intéresse ici à $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$ et on suppose qu'il existe $\beta \in \mathbb{R}^{p+1}$ et des VAs $\epsilon_1, \dots, \epsilon_n$ telles que:

$$y_i = \beta_0 + \beta_1 x_i + \dots + \beta_p x_i^p + \epsilon_i.$$

On peut alors réécrire: $Y_n = X_n \beta + \epsilon_n$. L'estimateur de carré minimal (OLS) de β est donné par: $\min_{\beta} \|Y_n - X_n \beta\|^2$.

Si on suppose par ailleurs que $\mathbb{E}(\epsilon_n) = 0$ et $Cov(\epsilon_n) = \sigma^2 I_n$, alors l'OLS est non biaisé et $Cov(\beta) = \sigma^2 (X_n^T X_n)^{-1}$ et, si $n > p + 1$, alors un estimateur de σ^2 est donné par:

$$\hat{\sigma}^2 = \frac{\|Y_n - X_n \hat{\beta}\|^2}{n - p - 1}$$

avec $\hat{\beta} = (X_n^T X_n)^{-1} X_n^T Y_n$ qui défini alors l'OLS.

Ensuite, si on suppose que $\epsilon_1, \dots, \epsilon_n$ sont des VAs gaussiennes indépendantes centrées de variance σ^2 , alors:

Proposition: Les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants et:

$$\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2 (X_n^T X_n)^{-1}) \text{ et } (n-p-1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_2(n-p-1).$$

4 Intervalles de Confiance

4.1 Définitions Générales

Soit $\alpha \in (0, 1/2)$ la précision désirée pour notre intervalle de confiance.

Définition: Intervalle de confiance

Un intervalle de confiance de niveau $1 - \alpha$ pour la QI $g(\theta)$ est un intervalle $I_n = [I_n^-, I_n^+]$ tel que I_n^- et I_n^+ soient des statistiques et, pour tout $\theta \in \Theta$, $\mathbb{P}_\theta(g(\theta) \in I_n) = 1 - \alpha$.

Il peut être assez difficile, voire impossible, de construire des intervalles de confiance comme définis précédemment (on dit qu'ils sont exacts), on introduit donc les définitions suivantes:

Définition:

Soit un intervalle I_n tel que I_n^- et I_n^+ soient des statistiques, alors cet intervalle est dit:

- de confiance asymptotique si, pour tout $\theta \in \Theta$, $\lim_n \mathbb{P}_\theta(g(\theta) \in I_n) = 1 - \alpha$
- de confiance par excès si, pour tout $\theta \in \Theta$, $\lim_n \mathbb{P}_\theta(g(\theta) \in I_n) \geq 1 - \alpha$.

4.2 Construction d'Intervalles de Confiance Exact

4.2.1 Fonction Pivot

On commence par une définition générale:

Définition: VA libre

Un VA Q est dite libre selon \mathbb{P}_θ si sa loi ne dépend pas de θ .

Définition: Fonction pivot

Un fonction pivot pour $g(\theta)$ est une fonction $\pi_n : E^n \times g(\Theta) \rightarrow \mathbb{R}$ telle que $\pi_n(X_n, g(\theta))$ est libre.

4.2.2 Exemple dans le Modèle Gaussien

On s'intéresse à la moyenne μ dans un modèle Gaussien qu'on estime classiquement avec \bar{X}_n . La loi de \bar{X}_n selon $\mathbb{P}_{\mu, \sigma^2}$ est $\mathcal{N}(\mu, \sigma^2/n)$. Il vient alors:

$$Y_n = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1) \text{ est libre.}$$

Néanmoins, la fonction:

$$\pi_n(x_n, \mu) = \frac{\bar{x}_n - \mu}{\sqrt{\sigma^2/n}}$$

n'est pas une fonction pivot puisqu'elle dépend de (μ, σ^2) à travers μ et σ et non uniquement de μ .

Supposons donc momentanément que σ^2 soit connu, alors π_n devient une fonction pivot. Il vient alors, pour tout réels a et b tels que $a < b$:

$$\mathbb{P}_{\mu, \sigma^2}(Y_n \in [a, b]) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp(-x^2/2) dx$$

donc, pour tout choix de a et b tels que:

$$\frac{1}{\sqrt{2\pi}} \int_a^b \exp(-x^2/2) dx = 1 - \alpha$$

alors l'intervalle $[\bar{X}_n - b\sqrt{\sigma^2/n}, \bar{X}_n - a\sqrt{\sigma^2/n}]$ est un intervalle de confiance exact de précision $1 - \alpha$ pour μ .

Rappelons par ailleurs que l'on définit par ϕ_r le quantile d'ordre r de la distribution gaussienne standard, alors a et b permettent de satisfaire l'égalité attendue si et seulement si:

$$\exists r \in [0, \alpha] : a = \phi_r, b = \phi_{r+1-\alpha}.$$

Pour une telle paire (a, b) et l'intervalle de confiance exact $[\bar{X}_n - b\sqrt{\sigma^2/n}, \bar{X}_n - a\sqrt{\sigma^2/n}]$, la probabilité de sous-estimer μ est r , celle de la sur-estimer est $\alpha - r$.

Supposons désormais que σ^2 ne soit pas connu. Une idée classique consiste alors à remplacer σ^2 par l'estimateur non-biaisé de la variance S_n^2 et on considère:

$$Y'_n = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \sim t(n-1) \text{ est libre.}$$

Il vient alors que:

$$\pi_n(x_n, \mu) = \frac{\bar{x}_n - \mu}{\sqrt{s_n^2/n}} \text{ avec } s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_i)^2$$

est une fonction pivot.

En conséquence, pour tout réels a et b tels que $a < b$, il vient:

$$\mathbb{P}_{\mu, \sigma^2}(Y'_n \in [a, b]) = \int_a^b p_{n-1}(x) dx \text{ avec } p_{n-1} \text{ la densité de la loi } t(n-1).$$

Une nouvelle fois, dès que le couple (a, b) vérifie:

$$\int_a^b p_{n-1}(x) dx = 1 - \alpha$$

on obtient un intervalle de confiance de précision $1 - \alpha$ pour μ .

4.2.3 Résumé de la Méthode

On commence par trouver une fonction pivot $Q_n = \pi_n(x_n, g(\theta))$, on réécrit la condition $Q_n \in [a, b]$ comme $g(\theta) \in I_n$ où les extrémités de I_n sont des statistiques. Enfin on choisit un couple (a, b) qui satisfait: $\mathbb{P}(Q_n \in [a, b]) = 1 - \alpha$, ce qui revient à choisir $a = q_{n,r}$ et $b = q_{n,r+1-\alpha}$ avec $0 \leq r \leq \alpha$ et $q_{n,r}$ le quantile d'ordre r de Q_n .

4.3 Construction d'Intervalle de Confiance Asymptotique

On rappelle que ϕ_r définit le quantile d'ordre r de la distribution gaussienne standard.

Proposition: Intervalle de confiance asymptotique

Soit Z_n un estimateur consistant et convergent normalement de $g(\theta)$, on note $V(\theta)$ sa variance asymptotique. On suppose qu'un estimateur consistant \hat{V}_n de $V(\theta)$ est connu, alors:

$$\forall \alpha \in (0, 1/2), I_n = [Z_n - \phi_{1-\alpha/2} \sqrt{\frac{\hat{V}_n}{n}}, Z_n + \phi_{1-\alpha/2} \sqrt{\frac{\hat{V}_n}{n}}]$$

est un intervalle de confiance asymptotique avec une précision de $1 - \alpha$ pour $g(\theta)$.

En général il n'est pas difficile de trouver un estimateur consistant de $V(\theta)$, dès que V est continue et que $\hat{\theta}_n$ est un estimateur consistant de θ , alors on a simplement: $\hat{V}_n = V(\hat{\theta}_n)$.

4.4 Construction d'Intervalle de Confiance par Excès

On s'intéresse au cas où on ne possède pas de fonction pivot, comme dans le cas d'une loi de Bernoulli. On va alors construire des intervalles de confiance par excès à l'aide des inégalités de concentration.

Définition: Inégalité de concentration

Une inégalité de concentration pour une variable aléatoire Y est une inégalité de la forme: $\mathbb{P}(|Y - \mathbb{E}(Y)| \geq r) \leq c_Y(r)$ pour une fonction de concentration c_Y convergente asymptotiquement vers 0.

Si Z_n est un estimateur non-biaisé de $g(\theta)$ et qui vérifie une inégalité de concentration telle que:

$$\forall r > 0, \sup_{\theta} \mathbb{P}_{\theta}(|Z_n - g(\theta)| \geq r) \leq c_{Z_n}(r)$$

alors tout $r_{n,\alpha} > 0$ tel que $c_{Z_n}(r_{n,\alpha}) \leq \alpha$ produit l'intervalle de confiance par excès: $[Z_n - r_{n,\alpha}, Z_n + r_{n,\alpha}]$ pour $g(\theta)$.

4.4.1 L'Inégalité de Bienaymé-Chebychev

C'est l'inégalité classique:

$$\mathbb{P}(|Y - \mathbb{E}(Y)| \geq a) \leq \frac{V(Y)}{a^2}.$$

Pour un modèle de Benoulli dans lequel on utilise \bar{X}_n comme estimateur de p , alors, pour tout $a > 0$, il vient:

$$\mathbb{P}_p(|\bar{X}_n - p| \geq a) \leq \frac{p(1-p)}{a^2}$$

donc, pour a tel que: $\frac{p(1-p)}{a^2} \leq \alpha$, il vient:

$$\mathbb{P}_p(p \in [\bar{X}_n - a/\sqrt{n}, \bar{X}_n + a/\sqrt{n}]) \leq 1 - \alpha.$$

Il reste à trouver une telle valeur de a qui ne dépende pas de p , pour les VAs bornées, on peut utiliser le lemme suivant:

Lemme: Borne universelle de la variance

Soit Y une VA à valeurs dans $[0, 1]$, alors: $V(Y) \leq \frac{1}{4}$.

Il suffit alors de prendre $a = \frac{1}{2\sqrt{\alpha}}$.

4.4.2 Inégalité de Hoeffding

On commence par introduire le lemme suivant:

Lemme: Inégalité de Hoeffding

Soient X_1, \dots, X_n des VAs iid à valeurs dans $[0, 1]$, alors, pour tout $n \geq 1$ et $r > 0$, il vient:

$$\mathbb{P}\left(\sum_{i=1}^n (X_i - \mathbb{E}(X_i)) \geq r\sqrt{n}\right) \leq \exp(-2r^2).$$

Il vient alors:

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| \geq r/\sqrt{n}) \leq 2 \exp(-2r^2).$$

5 Estimateur du Maximum de Vraisemblance

Jusqu'à présent, on a vu des méthodes d'estimation, de calcul d'intervalle de confiance qui reposent sur le choix de fonctions particulières. A l'inverse l'estimateur du maximum de vraisemblance permet d'estimer les paramètres d'un modèle de façon plus générale.

5.1 Vraisemblance d'un Echantillon et Estimateur

On s'intéresse à (X_1, \dots, X_n) des VAs iid dans \mathbb{R}^n .

Définition: Vraisemblance d'une observation

Soit $x_n^* = (x_1, \dots, x_n)$ une valeur possible de $X_n^* = (X_1, \dots, X_n)$, la vraisemblance de cette observation est donnée par la fonction:

$$L_n(x_n^*, \cdot) = \theta \mapsto \prod_{i=1}^n p(x_i, \theta) \quad \text{avec } p(x_i, \theta) = \mathbb{P}_\theta(X_i = x_i).$$

Dans le cas discret, alors: $L_n(x_n^*, \theta) = \mathbb{P}_\theta(X_1 = x_1, \dots, X_n = x_n)$, si \mathbb{P}_θ admet une densité selon la mesure de Lebesgue, alors elle est donnée par: $x_n^* \mapsto L_n(x_n^*, \theta)$.

L'estimation du maximum de vraisemblance consiste à chercher le paramètre θ^* qui rend les valeurs observées les plus probables.

Définition: Estimateur du maximum de vraisemblance

Supposons que, pour tout $x_n^*, \theta \mapsto L_n(x_n^*, \theta)$ atteigne un maximum global en $\theta^* = \theta_n(x_n^*)$. L'estimateur du maximum de vraisemblance (MLE) de θ est la statistique:

$$\hat{\theta}_n = \theta_n(x_n^*).$$

Dans le cas où il y a plusieurs maximum, on peut simplement prendre $\hat{\theta}_n \in \arg \max_\theta L_n(x_n^*, \theta)$.

Si la fonction de vraisemblance est différentiable, on peut alors calculer $\hat{\theta}_n$ en cherchant les zéros du gradient. Dans cette perspective, il peut être intéressant de considérer la dérivée du logarithme de la vraisemblance:

$$l_n(x_n^*, \theta) = \log(L_n(x_n^*, \theta)).$$

5.2 Exemples

5.2.1 Modèle de Bernoulli

On a alors:

$$L_n(x_n^*, p) = \prod_{i=1}^n \mathbb{P}_p(X_i = x_i) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$$

donc:

$$l_n(x_n^*, p) = \sum_{i=1}^n (x_i \log(p) + (1-x_i) \log(1-p)) = n\bar{X}_n \log(p) + n(1-\bar{X}_n) \log(1-p).$$

On peut alors calculer:

$$\frac{\partial l}{\partial p}(x_n^*, p^*) = 0 \iff (1-p^*)\bar{X}_n - p^*(1-\bar{X}_n) = 0 \iff p^* = \bar{X}_n.$$

Il reste alors à s'assurer que p^* correspond bien à un maximum, dans le cas du modèle de Bernoulli c'est bien le cas.

5.2.2 Cas du Modèle Gaussien

On a alors:

$$L_n(x_n^*, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}^n} \prod_{i=1}^n \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

donc:

$$\frac{\partial l_n}{\partial \mu}(x_n^*, \mu^*, \sigma^{2*}) = \frac{1}{\sigma^{2*}} \sum_{i=1}^n (x_i - \mu^*) \quad \text{et} \quad \frac{\partial l_n}{\partial \sigma^2}(x_n^*, \mu^*, \sigma^{2*}) = -\frac{n}{2\pi\sigma^*} + \sum_{i=1}^n \frac{(x_i - \mu^*)^2}{\sigma^{3*}}.$$

On trouve alors : $\mu^* = \bar{X}_n$ et $\sigma^{2*} = S_n^2$ qui sont les estimateurs qu'on trouve également avec la méthode des moments.

5.3 Optimalité du MLE

Lorsqu'on a plusieurs estimateurs du même paramètre on peut comparer leur MSE, voire leur variance asymptotique s'ils sont asymptotiquement normaux. Le MLE est asymptotiquement normal, sa variance asymptotique est par ailleurs optimale dans les modèles réguliers.

5.3.1 Modèle Régulier et Information de Fisher

On rappelle que $\nabla\phi_\theta = (\frac{\partial\phi}{\partial\theta_1}, \dots, \frac{\partial\phi}{\partial\theta_d})$.

Définition: Modèle régulier

Un modèle paramétrique est régulier si: Θ est ouvert et:

- $\forall x_1, \forall \theta \in \Theta, L_1(x_1, \theta) > 0$
- $\forall x_1, \theta \mapsto l_1(x_1, \theta) \in C^1(\Theta)$
- $\forall \theta \in \Theta, \mathbb{E}_\theta(||\nabla_\theta l_1(X_1, \theta)||^2) < +\infty$.

Définition: Score dans un modèle régulier

Dans un modèle régulier, on appelle score le vecteur aléatoire $\nabla_\theta l_1(X_1, \theta) = (\frac{1}{p(X_1, \theta)} \frac{\partial p}{\partial \theta_i}(X_1, \theta))_i$.

L'information de Fisher $I(\theta)$ d'un modèle régulier est:

$$I(\theta) = Cov_\theta(\nabla_\theta l_1(X_1, \theta)).$$

Proposition:

On a, pour tout θ , $\mathbb{E}_\theta(\nabla_\theta l_1(X_1, \theta)) = 0$ et les coefficients de $I(\theta)$ sont:

$$I_{i,j}(\theta) = \mathbb{E}_\theta\left(\frac{\partial l_1}{\partial \theta_i}(X_1, \theta) \frac{\partial l_1}{\partial \theta_j}(X_1, \theta)\right) = -\mathbb{E}_\theta\left(\frac{\partial^2 l_1}{\partial \theta_i \partial \theta_j}(X_1, \theta)\right).$$

5.3.2 Estimateur Efficace

On peut utiliser l'information de Fisher pour définir une notion d'efficacité pour les modèles non-biaisés.

Théorème: Borne de Cramér-Rao

Pour un modèle régulier tel que $I(\theta) \geq 0$ pour tout θ , on note $\tilde{\theta}_n = t_n(X_n)$ un estimateur non-biaisé de θ avec la matrice de covariance $K_n(\theta) = Cov_\theta(\tilde{\theta}_n)$; alors:

$$K_n(\theta) \succeq \frac{I^{-1}(\theta)}{n}.$$

Définition: Estimateur efficace

Un estimateur est dit efficace s'il est non-biaisé et si sa matrice de covariance vérifie $K_n(\theta) = \frac{I^{-1}(\theta)}{n}$.

On rappelle que si $\tilde{\theta}_n$ est un estimateur non-biaisé, alors: $MSE(\tilde{\theta}_n, \theta) = \mathbb{V}_\theta(\tilde{\theta}_n) = tr(K_n(\theta)) \geq \frac{tr(I^{-1}(\theta))}{n}$. En conséquences, dans un modèle régulier, les estimateurs efficaces minimisent la MSE pour les estimateurs non-biaisés.