VECTEURS GAUSSIENS

Soit d un entier ≥ 1. On identifiera les éléments de $\mathbb{R}^d$ à des vecteurs colonnes.

1 Matrice de covariance

Soit X une variable aléatoire à valeurs dans $\mathbb{R}^d$. On note $X = {}^t(X_1 \ \cdots \ X_d)$. Les variables aléatoires réelles $X_1, \ldots, X_d$ sont appelées les *composantes* de X et leurs lois sont les *lois marginales* de la loi de X. On définit classiquement son espérance (sa moyenne)

$$m_X = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{pmatrix}$$

et aussi, si $X_1, \ldots, X_d$ sont de carré intégrable, sa variance

$$\begin{aligned} \Gamma_X = \mathrm{Var}(X) &= \mathbb{E}\big[(X - \mathbb{E}[X])^t(X - \mathbb{E}[X])\big] \\ &= (\mathrm{Cov}(X_i, X_j))_{1 \leq i,j \leq d} \\ &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathrm{Var}(X_d) \end{pmatrix}, \end{aligned}$$

où on rappelle que, pour toutes variables aléatoires réelles U et V de carré intégrable, $\mathrm{Cov}(U, V) = \mathbb{E}[(U - \mathbb{E}[U])(V - \mathbb{E}[V])] = \mathbb{E}[UV] - \mathbb{E}[U]\mathbb{E}[V]$ (covariance de U et V). La matrice Γ_X est la *matrice de covariance* de X (ou, plutôt, de la loi de X).
En particulier, si $X_1, \ldots, X_d$ sont indépendantes deux à deux, alors Γ_X est une matrice diagonale ayant pour coefficients diagonaux $\mathrm{Var}(X_1), \ldots, \mathrm{Var}(X_d)$.
Comme $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$ (ou comme ${}^t(A^tA) = A^tA$), la matrice Γ_X est symétrique. Elle est de plus positive : pour tout vecteur $u \in \mathbb{R}^d$,

$${}^tu\Gamma_X u = \sum_{i,j} u_i \,\mathrm{Cov}(X_i, X_j) u_j = \sum_{i,j} \mathrm{Cov}(u_i X_i, u_j X_j) = \mathrm{Var}(u_1 X_1 + \cdots + u_d X_d) \geq 0.$$

(En revenant aux définitions de Cov et Var, la dernière égalité est simplement le développement du carré d'une somme)
Au passage, on voit que $\mathrm{Var}({}^tuX) = {}^tu\Gamma_X u$. Plus généralement, pour toute matrix A de taille (p, d), où $p \geq 1$, le vecteur AX est une variable aléatoire à valeurs dans $\mathbb{R}^p$, sa moyenne est Am_X (par linéarité de l'espérance) et sa variance est

$$\begin{aligned} \Gamma_{AX} &= \mathbb{E}\big[(AX - Am)^t(AX - Am)\big] = \mathbb{E}\big[A(X - m)^t(X - m)^tA\big] \\ &= A\mathbb{E}\big[(X - m)^t(X - m)\big]^tA = A\Gamma_X{}^tA. \end{aligned}$$

En particulier (avec $p = d$), si $\Gamma_X = I_d$ (par exemple si $X_1, \ldots, X_d$ sont indépendants et de variance 1), alors $\Gamma_{AX} = A^tA$. Or toute matrice symétrique positive Γ peut s'écrire $\Gamma = A^tA$: par exemple, partant de la diagonalisation $\Gamma = PD^tP$ en base orthonormale où $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ avec $\lambda_1 \geq 0, \ldots, \lambda_d \geq 0$, on peut prendre $A = P\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})$. On a alors $\Gamma_{AX} = A^tA = \Gamma$ si $\Gamma_X = I_d$ (Remarque : une autre matrice A possible est obtenue par la méthode de Cholesky). Ceci montre que les matrices de covariance sont les matrices symétriques positives.
Le rang de la matrice de covariance a une interprétation importante :

Proposition 1. Γ_X est de rang $\leq r$ si, et seulement s'il existe un sous-espace affine V de dimension r tel que $X \in V$ p.s.

Démonstration. On a vu que, pour tout $u \in \mathbb{R}^d$, $\mathrm{Var}(\langle u, X\rangle) = {}^t u \Gamma_X u$. Et une variable aléatoire de carré intégrable est constante p.s. (égale à sa moyenne p.s.) si, et seulement si sa variance est nulle. En particulier, ${}^t u \Gamma_X u = 0$ si, et seulement si $\langle u, X\rangle = \langle u, m\rangle$ p.s., ce qui équivaut à dire que $X - m \perp u$ p.s.. En appliquant ceci à toute base orthogonale de $\mathbb{R}^d$ qui commence par une base de $\ker \Gamma_X$, on obtient que

$$X \in m + \big(\ker \Gamma_X\big)^\perp \quad \text{p.s..}$$

($\ker \Gamma = \{x \in \mathbb{R}^d | \Gamma x = 0\} = \{x \in \mathbb{R}^d | {}^t x \Gamma x = 0\}$ si Γ est symétrique positive) et que $m + (\ker \Gamma_X)^\perp$ est le plus petit sous-espace pour lequel c'est vérifié, d'où la proposition. □

2 Vecteurs gaussiens

Soit $Z = {}^t\big(Z_1 \cdots Z_d\big)$, où $Z_1, \ldots, Z_d$ sont des variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$, m un vecteur de $\mathbb{R}^d$, et Γ une matrice symétrique positive de taille d. On choisit une matrice carrée A telle que $A {}^t A = \Gamma$. Par ce qui précède, $X = m + AZ$ est un vecteur aléatoire tel que

$$m_X = m \qquad \text{et} \qquad \Gamma_X = A \Gamma_Z {}^t A = A I_d {}^t A = \Gamma.$$

Calculons la fonction caractéristique de X. Pour tout $u \in \mathbb{R}^d$, on a

$${}^t u X = {}^t u m + {}^t u A Z = {}^t u m + {}^t({}^t A u) Z$$

donc (en notant $\langle \cdot, \cdot \rangle$ le produit scalaire usuel de $\mathbb{R}^d$ pour simplifier les notations)

$$\Phi_X(u) = \mathbb{E}[e^{i {}^t u X}] = e^{i\langle u, m\rangle} \Phi_Z({}^t A u).$$

Les composantes de Z sont indépendantes et de loi $\mathcal{N}(0,1)$ donc

$$\Phi_Z(u) = \Phi_{X_1}(u_1) \cdots \Phi_{X_d}(u_d) = e^{-\frac{1}{2}u_1^2} \cdots e^{-\frac{1}{2}u_d^2} = e^{-\frac{1}{2}\|u\|^2} = e^{-\frac{1}{2}{}^t u u}.$$

On en déduit

$$\Phi_X(u) = e^{i\langle u, m\rangle} e^{-\frac{1}{2}{}^t u A {}^t A u} = e^{i\langle u, m\rangle} e^{-\frac{1}{2}{}^t u \Gamma u}.$$

On remarque notamment que la dernière expression ne dépend pas du choix de A, donc la loi de X non plus. $X = m + AZ$ est appelé un *vecteur gaussien* de moyenne m et de covariance Γ et sa loi est appelée *loi gaussienne de moyenne m et de covariance* Γ, abrégée en $\mathcal{N}(m, \Gamma)$. La loi de Z, à savoir $\mathcal{N}(0, I_d)$ est appelée *loi gaussienne standard de* $\mathbb{R}^d$.
La proposition suivante caractérise les vecteurs gaussiens, c'est-à-dire suivant une loi $\mathcal{N}(m, \Gamma)$ pour certain vecteur m et un certaine matrice symétrique positive Γ :

Proposition 2. Soit X un vecteur aléatoire de $\mathbb{R}^d$. X est un vecteur gaussien si, et seulement si toute combinaison linéaire de composantes de X suit une loi gaussienne (sur $\mathbb{R}$).

Démonstration. Une combinaison linéaire des composantes de X s'écrit $\langle a, X\rangle = {}^t a X$ où $a \in \mathbb{R}^d$. Si m et Γ sont l'espérance et la variance de X (qui existent sous chacune des deux hypothèses), alors $\mathbb{E}[\langle a, X\rangle] = \langle a, m\rangle$ et $\mathrm{Var}(\langle a, X\rangle) = {}^t a \Gamma a$ (vu dans la partie 1).
Ainsi, $\langle a, X\rangle$ suit une loi gaussienne pour tout $a \in \mathbb{R}^d$ si, et seulement si, pour tout $a \in \mathbb{R}^d$ sa fonction caractéristique est

$$\Phi_{\langle a, X\rangle}(u) = e^{iu\langle a, m\rangle} e^{-\frac{1}{2}({}^t a \Gamma a)u^2} = e^{i\langle ua, m\rangle} e^{-\frac{1}{2}{}^t(ua)\Gamma(ua)}$$

et X est un vecteur gaussien si, et seulement si sa fonction caractéristique est

$$\Phi_X(a) = e^{i\langle a, m\rangle} e^{-\frac{1}{2}{}^t a \Gamma a}.$$

Comme, de façon générale, $\Phi_{\langle a, X\rangle}(u) = \Phi_X(ua)$, l'équivalence entre les deux propriétés précédentes se déduit immédiatement, d'où la conclusion. □

Cette proposition permet de définir des vecteurs gaussiens dans un espace de dimension infinie.
La définition implique que si $X \sim \mathcal{N}(m, \Gamma)$ et si A est une matrice de taille (p, d) et $b \in \mathbb{R}^d$, alors $AX + b \sim \mathcal{N}(am + b, A\Gamma^t A)$. (La définition montre que c'est un vecteur gaussien et ses paramètres se calculent comme dans la première partie)
En particulier, si Z est un vecteur gaussien standard et P est une matrice orthogonale ($P^tP = I_d$), alors PZ est encore un vecteur gaussien standard. Autrement dit, la loi gaussienne standard (et plus généralement $\mathcal{N}(0, \sigma^2 I_d)$) est invariante par les rotations de centre O (et par les symétries orthogonales d'axe passant par l'origine). Cette propriété se voit aussi sur la densité de la loi : elle est radiale.
Si $Z \sim \mathcal{N}(0, I_d)$, on sait que Z a pour densité $z \mapsto \frac{1}{\sqrt{2\pi}} e^{-z_1^2} \cdots \frac{1}{\sqrt{2\pi}} e^{-z_d^2} = (2\pi)^{-d/2} e^{-\frac{1}{2}\|z\|^2}$ puisque les composantes sont indépendantes. Pour le vecteur gaussien $X = m + AZ \sim \mathcal{N}(m, \Gamma)$ (où $\Gamma = A^tA$), la proposition 1 montre que si Γ n'est pas inversible, X ne peut pas avoir de densité. En revanche, si Γ est inversible, un changement de variable montre que X a pour densité

$$f_X : x \mapsto \frac{1}{\sqrt{(2\pi)^d \det \Gamma}} e^{-\frac{1}{2}{}^t(x-m)\Gamma^{-1}(x-m)}.$$

Si Γ est diagonale, alors les composantes $X_1, \ldots, X_d$ d'un vecteur $X \sim \mathcal{N}(m, \Gamma)$ sont indépendantes. Cela se voit sur la densité ci-dessus. Ou simplement parce que dans ce cas on peut prendre $A = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ d'où $X_i = m_i + \sigma_i Z_i$. La réciproque est toujours vraie. Ainsi, les composantes d'un vecteur gaussien sont indépendantes si, et seulement si sa matrice de covariance est diagonale. Le théorème de Cochran généralise cette remarque.

Théorème 1 – *Théorème de Cochran.* Soit $X \sim \mathcal{N}(m, \sigma^2 I_d)$, et $\mathbb{R}^d = E_1 \overset{\perp}{\oplus} \ldots \overset{\perp}{\oplus} E_r$ une décomposition de $\mathbb{R}^d$ en sous-espaces (affines) orthogonaux. Pour $k = 1, \ldots, r$, on note d_k la dimension de E_k, et P_k la projection orthogonale sur E_k. Alors les variables aléatoires $Y_1 = P_1X_1, \ldots, Y_r = P_rX$ sont indépendantes, et

$$Y_k \sim \mathcal{N}(P_k m, \sigma^2 P_k) \quad \text{et} \quad \frac{1}{\sigma^2}\|Y_k - P_k m\|^2 \sim \chi^2_{d_k}.$$

Dans l'énoncé, χ^2_d est la loi du χ^2 (khi-deux) à d degrés de liberté, c'est-à-dire par définition la loi de la variable aléatoire $Z_1^2 + \cdots + Z_r^2$ où $Z_1, \ldots, Z_r$ sont i.i.d. de loi $\mathcal{N}(0, 1)$.

Démonstration. Il suffit de montrer l'indépendance. La suite en résulte (avec ce qui précède, et le fait que $P^tP = P^2 = P$ pour toute projection orthogonale P).
Pour $k = 1, \ldots, r$, on choisit une base orthonormale $(e_{k,1}, \ldots, e_{k,d_k})$ de E_k, de sorte que $(e_1, \ldots, e_d) = (e_{1,1}, \ldots, e_{1,d_1}, \ldots, e_{r,1}, \ldots, e_{r,d_r})$ est une base orthonormale de $\mathbb{R}^d$. Les variables aléatoires $\langle X, e_i \rangle$, $i = 1, \ldots, d$, sont les composantes de X dans cette base. Or la loi de X est invariante par les applications orthogonales (vu plus haut) donc en particulier par les changements de base entre bases orthonormales. Par suite, les variables aléatoires $\langle X, e_i \rangle$, $i = 1, \ldots, d$, sont i.i.d. de loi $\mathcal{N}(\langle m, e_i \rangle, \sigma^2)$.
Comme $Y_k = P_kX = A_k + \langle X, e_{k,1} \rangle e_{k,1} + \cdots + \langle X, e_{k,d_k} \rangle e_{k,d_k}$, où $A_k = P_k m$, l'indépendance annoncée résulte de la propriété d'« indépendance par paquets » : $Y_1, \ldots, Y_r$ dépendent respectivement de paquets de variables disjoints parmi une famille de variables indépendantes, donc sont indépendantes entre elles. □

Remarque. On n'a donné qu'un cas particulier du théorème de Cochran, qui est le cas usuel. L'énoncé général (dont la preuve est essentiellement identique) suppose $X \sim \mathcal{N}(m, \Gamma)$ où Γ est inversible et $\mathbb{R}^d = E_1 \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} E_r$ au sens du produit scalaire $(u, v) \mapsto \langle u, v \rangle_\Gamma = {}^t u \Gamma v$, et conclut que les projections orthogonales (au sens de ce même produit scalaire) de X sur $E_1, \ldots, E_r$ sont indépendantes. Les variables $\|Y_k - P_k m\|^2$ ne suivent cependant pas des lois du χ^2 (ce sont des sommes de variables $\mathcal{N}(0, 1)$ pondérées par les valeurs propres des différents sous-espaces), et la variance de Y_k est $P_k \Gamma^t P_k$.

Enfin, les lois gaussiennes sur $\mathbb{R}^d$ interviennent dans la généralisation du théorème central limite aux variables aléatoires à valeurs dans $\mathbb{R}^d$:

Théorème 2 – *Théorème central limite multidimensionnel.* Soit $X_1, X_2, \ldots$ une suite de vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, i.i.d., de carré intégrable, de moyenne m et de variance Γ. Alors

$$\frac{1}{\sqrt{n}} (X_1 + X_2 + \cdots + X_n - nm) \xrightarrow[n]{\text{(loi)}} \mathcal{N}(0, \Gamma).$$

Démonstration. Même preuve que la version réelle. En remplaçant X_i par $X_i - m$ on se ramène à $m = 0$. Soit $u \in \mathbb{R}^d$. On a

$$\Phi_{\frac{1}{\sqrt{n}}(X_1+X_2+\cdots+X_n)}(u) = \Phi_X\left(\frac{u}{\sqrt{n}}\right)^n$$

(où $\Phi_X = \Phi_{X_1}$). On a le développement limité

$$\Phi_X(t) = 1 - \frac{1}{2}{}^t t\Gamma t + \underset{t\to 0}{o}\left(\|t\|^2\right),$$

d'où

$$\Phi_{\frac{1}{\sqrt{n}}(X_1+X_2+\cdots+X_n)}(u) = \left(1 - \frac{1}{2n}{}^t u\Gamma u + \underset{n}{o}\left(\frac{1}{n}\right)\right)^n \underset{n}{\longrightarrow} \exp\left(-\frac{1}{2}{}^t u\Gamma u\right).$$

NB : On a utilisé $(1 - z_n/n)^n = \exp(n\,\mathrm{Ln}(1 - z_n/n)) \to_n \exp(-c)$ où $(z_n)_n$ est une suite *complexe* telle que $z_n \to_n c$ et Ln est la détermination principale du logarithme complexe. L'égalité est vraie dès que $|1 - z_n/n| < 1$, donc pour n grand, et la limite vient de $\mathrm{Ln}(1 - z) \sim -z$, quand $z \to 0$, $z \in \mathbb{C}$. (Si on définit $z \mapsto \mathrm{Ln}(1 - z) = -\sum_{k=1}^{\infty} \frac{1}{k} z^k$ sur $D(0, 1)$, c'est une série entière qui coïncide avec $x \mapsto \ln(1 - x)$ sur $]0, 1[$; elle vérifie donc $\exp(\mathrm{Ln}(1 - z)) = 1 - z$ pour $z \in]0, 1[$ et, par unicité du prolongement analytique, pour tout $z \in D(0, 1)$; et $\mathrm{Ln}(1 - z) \sim -z$ quand $z \to 0$ vu le premier terme du développement en série entière.) □

NOTIONS DE STATISTIQUES

Rappels sur les vecteurs gaussiens

Soit $d \in \mathbb{N}^*$. Un **vecteur gaussien** est un vecteur aléatoire à valeurs dans $\mathbb{R}^d$ dont toutes les combinaisons affines des composantes suivent des lois normales. La loi d'un tel vecteur aléatoire est caractérisée par sa moyenne m et sa matrice de covariance Γ, on la note $\mathcal{N}(m, \Gamma)$.
Pour toute matrice A de taille (p, d) et tout $b \in \mathbb{R}^p$,

$$\text{si } X \sim \mathcal{N}(m, \Gamma) \text{ alors } AX + b \sim \mathcal{N}(Am + b, A\Gamma {}^tA).$$

En particulier, si $\Gamma = A^tA$ (A est appelée une racine carrée de Γ), et si $X \sim \mathcal{N}(0, I_d)$ alors $m + AX \sim \mathcal{N}(m, \Gamma)$. Ceci permet de toujours se ramener au cas $\mathcal{N}(0, I_d)$, la *loi gaussienne standard de* $\mathbb{R}^d$, où les composantes sont indépendantes et de loi $\mathcal{N}(0, 1)$.
Autre conséquence, si $X \sim \mathcal{N}(0, \sigma^2 I_d)$ et A est une matrice orthogonale (rotation, symétrie orthogonale), $A^tA = I_d$ donc AX a même loi que X. Le théorème suivant s'en déduit.

Théorème 3 – ***Théorème de Cochran.*** Soit $X \sim \mathcal{N}(m, \sigma^2 I_d)$, et $\mathbb{R}^d = E_1 \overset{\perp}{\oplus} \dots \overset{\perp}{\oplus} E_r$ une décomposition de $\mathbb{R}^d$ en sous-espaces (affines) orthogonaux. Pour $k = 1, \dots, r$, on note d_k la dimension de E_k, et P_k la projection orthogonale sur E_k. Alors les variables aléatoires $Y_1 = P_1X_1, \dots, Y_r = P_rX$ sont indépendantes,

$$Y_k \sim \mathcal{N}(P_k m, \sigma^2 P_k) \quad \text{et} \quad \frac{1}{\sigma^2}\|Y_k - P_k m\|^2 \sim \chi^2_{d_k}.$$

(voir à la fin pour la définition de la loi $\chi^2_{d_k}$)

Démonstration. En remplaçant X par $X - m$, on peut supposer que $m = 0$.
Pour $k = 1, \dots, r$, on note $(f_{k,1}, \dots, f_{k,d_k})$ une base orthonormale de E_k, de sorte que leur concaténation fournit une base orthonormale $(f) = (f_1, \dots, f_d)$ de $\mathbb{R}^d$. Le vecteur

$$\begin{pmatrix} \langle X, f_1 \rangle \\ \vdots \\ \langle X, f_d \rangle \end{pmatrix}$$

des composantes de X dans la base (f) est l'image de X par une application orthogonale (l'application de changement de base A définie par $A(f_i) = e_i$ où $(e_1, \dots, e_d)$ est la base canonique de $\mathbb{R}^d$). Par conséquent, ce vecteur a même loi que X, c'est-à-dire que ses composantes sont indépendantes et suivent la loi $\mathcal{N}(0, \sigma^2)$. Or chacune des projections

$$Y_k = P_kX = \langle X, f_{k,1} \rangle f_{k,1} + \dots + \langle X, f_{k,d_k} \rangle f_{k,d_k}$$

pour $k = 1, \dots, r$, ne dépend que d'une sous-famille de composantes du vecteur précédent, et ces sous-familles sont disjointes les unes des autres, donc $Y_1, \dots, Y_r$ sont indépendantes (propriété d'indépendance « par paquets »).
De plus, on a $Y_k = P_kX \sim \mathcal{N}(0, \sigma^2 P_k {}^tP_k)$ et $P_k {}^tP_k = P_k^2 = P_k$ (projection orthogonale), et

$$\|Y_k\|_2 = |\langle X, f_{k,1} \rangle|^2 + \dots + |\langle X, f_{k,d_k} \rangle|^2$$

d'où l'on déduit la deuxième partie vu la définition de $\chi^2_{d_k}$. □

1 Principes généraux : estimation et test

L'objet des statistiques (plus exactement, des statistiques *mathématiques*, ou *inférentielles*) est de déduire de l'observation de données supposées aléatoires des informations sur la loi de celles-ci.
On se restreint dans la suite au cas où les données observées sont des réalisations de n variables aléatoires $X_1,\dots,X_n$ indépendantes et de même loi μ sur $\mathbb{R}^d$. On dit que $(X_1,\dots,X_n)$ est un échantillon de taille n.
On peut globalement distinguer deux types de problématiques :
– l'estimation d'une grandeur numérique dépendant de μ (son espérance, sa variance,...) : c'est une approche *quantitative*. Les notions concernées sont les *estimateurs* et les *régions de confiance*.
– confirmer ou infirmer des hypothèses sur la loi μ (Son espérance est-elle égale à 5 ? μ est-elle une loi de Poisson ? μ (sur $\mathbb{R}^d$) est-elle une loi produit ?...) : c'est une approche *qualitative*. Notons que cela revient à estimer des grandeurs discrètes dépendant de μ (autrement dit, des fonctions indicatrices). La notion qui s'y rattache est celle de *test d'hypothèse*.

1.1 Estimation

1.1.1 Estimateurs

Supposons que l'on souhaite estimer la valeur d'une quantité $\theta \in \mathbb{R}^d$ dépendant de μ. Typiquement, son espérance ou sa variance. Un **estimateur** de θ est une variable aléatoire $\widehat{\theta}$ à valeurs dans $\mathbb{R}^d$, qui dépend de $X_1,\dots,X_n$. Il est **consistant** (resp. **fortement consistant**) si $\widehat{\theta} \to \theta$ quand $n \to \infty$ en probabilité (resp. presque sûrement). La consistance est bien sûr une propriété essentielle pour un estimateur. On dit que $\widehat{\theta}$ est **sans biais** s'il est intégrable et $\mathbb{E}[\widehat{\theta}] = \theta$.

Exemple. Si μ a une espérance finie, on peut, pour estimer son espérance m, utiliser la **moyenne empirique**

$$\widehat{m} = \overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

C'est un estimateur sans biais (linéarité de l'espérance) et fortement consistant (loi forte des grands nombres).

Remarque. Bien que l'on se soit restreint ci-dessus à $\theta \in \mathbb{R}^d$, on peut considérer aussi des quantités vivant dans des espaces de dimension infinie, en adaptant alors la notion de consistance. Par exemple on peut tout simplement souhaiter estimer μ elle-même (où μ est une mesure sur $\mathbb{R}^d$). Pour cela, un choix naturel est la **mesure empirique**

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

C'est la mesure sur $\mathbb{R}^d$ telle que $\widehat{\mu}(A)$ est la proportion de points de l'échantillon qui appartiennent au borélien A. Par la loi forte des grands nombres (et la séparabilité de $\mathcal{C}_c(\mathbb{R}^d)$), c'est un estimateur fortement consistant, au sens où, presque sûrement $\widehat{\mu} \to \mu$ en loi quand $n \to \infty$.

Souvent, on suppose que μ appartient à un ensemble $\{\mu_\theta | \theta \in \Theta\}$ de mesures de probabilités indexé par $\Theta \subset \mathbb{R}^d$. Dans ce cadre (qui revient à dire que θ caractérise μ), θ apparaît comme un **paramètre** de la loi μ, et on parle d'*estimation paramétrique*. Le modèle est dit **identifiable** si θ est défini uniquement, c'est-à-dire que $\theta \mapsto \mu_\theta$ est injective sur Θ.
On peut souvent deviner une fonction qui constitue un (bon) estimateur de θ. Néanmoins, il y a aussi des méthodes plus ou moins générales :

Méthode des moments. Si $\theta = \varphi(m_1, m_2, \dots, m_k)$ est une fonction continue φ des k premiers moments de μ :

$$m_1 = m = \int x d\mu, \quad m_2 = \int x^2 d\mu, \quad \cdots \quad m_k = \int x^k d\mu,$$

alors, d'après la loi forte des grands nombres, un estimateur fortement consistant est

$$\widehat{\theta} = \varphi(\widehat{m}_1, \dots, \widehat{m}_k)$$

où

$$\widehat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \cdots \quad \widehat{m}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i)^k.$$

EMV. Si $\mu \in \{\mu_\theta | \theta \in \Theta\}$ et qu'il existe une mesure ν (pas nécessairement une probabilité) telle que $\mu_\theta \ll \nu$ pour tout $\theta \in \Theta$ (en général ν est la mesure de comptage, ou la mesure de Lebesgue), on appelle la densité

$$(x_1, \dots, x_n) \mapsto L_\theta(x_1, \dots, x_n) = \frac{d\mu_\theta^{\otimes n}}{d\nu^{\otimes n}}(x_1, \dots, x_n) = \frac{d\mu_\theta}{d\nu}(x_1) \cdots \frac{d\mu_\theta}{d\nu}(x_n)$$

la **vraisemblance** de θ. C'est une façon de mesurer la probabilité, sous μ_θ, d'observer un échantillon « voisin » de $(x_1, \dots, x_n)$. L'**estimateur du maximum de vraisemblance** est la valeur $\widehat{\theta}$ (pas nécessairement unique) telle que

$$L_{\widehat{\theta}}(X_1, \dots, X_n) = \sup_{\theta \in \Theta} L_\theta(X_1, \dots, X_n).$$

Cet estimateur n'est pas toujours simple à déterminer mais sa définition en fait un estimateur naturel et, sous des hypothèses de régularité de $\theta \mapsto L_\theta$ (que je ne précise pas), on peut montrer qu'il est consistant (et asymptotiquement gaussien).

Exemples d'EMV. a) On souhaite estimer la moyenne $\theta = \frac{1}{\lambda}$ de μ, où l'on sait que $\mu = \mathcal{E}(\lambda)$ (loi exponentielle) pour un certain $\lambda > 0$ inconnu. Notons $\mu_\theta = \mathcal{E}(1/\theta)$ pour suivre les notations ci-dessus. On prend $\nu = \mathbf{1}_{[0,\infty[}(x)dx$. Alors, pour $x > 0$,

$$\frac{d\mu_\theta}{d\nu}(x) = \frac{1}{\theta} e^{-x/\theta}$$

d'où

$$L_\theta(X_1, \dots, X_n) = \frac{1}{\theta^n} e^{-(X_1 + \cdots + X_n)/\theta}.$$

Comme souvent, il est plus pratique d'étudier la **log-vraisemblance** $\ln(L_\theta(X_1, \dots, X_n))$. Une étude de fonction donne immédiatement

$$\widehat{\theta} = \frac{n}{X_1 + \cdots + X_n}.$$

b) On souhaite estimer les paramètres $p = (p_1, \dots, p_r)$ d'une loi discrète $\mu = \mu_p$ sur $\{1, \dots, r\}$ (avec $\mu_p(\{i\}) = p_i$). On prend pour ν la mesure de comptage sur $\{1, \dots, r\}$ de sorte que

$$\frac{d\mu_p}{d\nu}(x) = \mu_p(\{x\}) = p_x$$

et donc

$$L_p(X_1, \dots, X_n) = p_{X_1} \cdots p_{X_n} = (p_1)^{N_1} \cdots (p_r)^{N_r}$$

où $N_k = \text{Card}\{1 \leq i \leq n | X_i = k\}$. On veut maximiser $\ln L_p = N_1 \ln p_1 + \cdots + N_r \ln p_r$, parmi les vecteurs p tels que $p_i \geq 0$ et $p_1 + \cdots + p_r = 1$. C'est un problème d'optimisation sous contrainte. On peut le résoudre en utilisant un multiplicateur de Lagrange. $p \mapsto \ln L_p$ tend vers $-\infty$ au bord du domaine donc son maximum est atteint à l'intérieur. Alors, il existe $\lambda \in \mathbb{R}$ tel que $p \mapsto \ln L_p - \lambda(p_1 + \cdots + p_r - 1)$ a un point critique en $\widehat{p}$. En dérivant par rapport à p_i on trouve que $N_i/\widehat{p}_i = \lambda$ d'où, via la contrainte, $\lambda = n$ et

$$\widehat{p}_i = \frac{N_i}{n}$$

(il y a un seul point critique donc par ce qui précède c'est $\widehat{p}$).

1.1.2 Régions de confiance

On souhaite en général disposer de bornes autour de la valeur approchée fournie par un estimateur, afin de mesurer l'erreur possible sur θ (θ est appelé la *valeur vraie*).
Soit $\alpha \in]0, 1[$ (en général, $\alpha = 5\%$ ou 1%). Une **région de confiance** de **niveau** $1 - \alpha$ est une partie $\mathcal{A}$ de $\mathbb{R}^d$ qui dépend de $X_1, \dots, X_n$ (donc « aléatoire ») telle que $\mathbb{P}(\theta \in \mathcal{A}) \geq 1 - \alpha$. Une **région de confiance asymptotique** de niveau $1 - \alpha$ est telle que $\lim \mathbb{P}(\theta \in \mathcal{A}) \geq 1 - \alpha$ quand $n \to \infty$.
Dans $\mathbb{R}$, on considère en général des régions de confiance qui sont des intervalles, et sont donc appelés **intervalles de confiance**.

Exemple. Si $X_1, \dots, X_n$ sont i.i.d. de carré intégrable, de moyenne m et de variance σ^2, un intervalle de confiance se déduit de l'inégalité de Tchebycheff : pour tout A,

$$\mathbb{P}\left(|\overline{X}_n - m| > A\right) \leq \frac{\sigma^2}{nA^2}$$

(car $\mathrm{Var}(\overline{X}_n) = \sigma^2/n$), donc l'intervalle

$$\left[\overline{X}_n - \sqrt{\frac{\sigma^2}{n\alpha}}, \overline{X}_n + \sqrt{\frac{\sigma^2}{n\alpha}}\right]$$

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

1.2 Tests

Les estimations peuvent suggérer certaines conclusions au sujet de la loi μ. Le rôle des tests est de décider si les données permettent effectivement, à certains niveaux d'erreur près, de tirer ces conclusions. Là où l'estimation donne une valeur numérique approchée de certaines grandeurs, le test produit un résultat binaire permettant éventuellement de prendre une décision.
Il convient d'être prudent. L'issue d'un test sera en général : « on rejette l'hypothèse » ou « on ne peut rejeter l'hypothèse », et on ne conclura pas, en revanche, que les données permettent d'accepter l'hypothèse. En cela, le choix de l'hypothèse est important et dépend des finalités de l'étude : il faut que le risque dont on souhaite minimiser la probabilité qu'il se produise soit celui de rejeter l'hypothèse à tort (*faux négatif*, ou *erreur de première espèce*). Par exemple, dans un test d'efficacité d'un médicament qui vient d'être mis au point, la sécurité sociale (qui, avec un budget en déficit, ne souhaite pas rembourser un médicament inefficace) peut vouloir tester l'hypothèse « le médicament est inefficace », quand le laboratoire pharmaceutique (qui souhaite rentabiliser son investissement en commercialisant le médicament) veut tester l'hypothèse « le médicament est efficace ».
Considérons deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ sur μ, autrement dit $\mathcal{H}_0$ et $\mathcal{H}_1$ sont des ensembles de lois auxquels on se demande si μ appartient. Le « test de $\mathcal{H}_0$ contre $\mathcal{H}_1$ » consiste à décider si, en fonction des données, on peut rejeter l'hypothèse $\mathcal{H}_0$ au profit de $\mathcal{H}_1$ ou ne pas la rejeter. On appelle $\mathcal{H}_0$ l'*hypothèse nulle*, c'est celle que l'on ne veut pas manquer de détecter. Ce test est de **niveau** α si, sous $\mathcal{H}_0$ (c'est-à-dire si $\mathcal{H}_0$ est vraie), le test est négatif avec probabilité $\leq \alpha$ (un niveau classique est 5%). Sa **puissance** $1-\beta$ est la probabilité que le test soit négatif sous $\mathcal{H}_1$ (où β est donc la probabilité de *faux positif*, ou *erreur de seconde espèce*). Souvent, on ne contrôle pas la puissance du test (d'où la dissymétrie entre $\mathcal{H}_0$ et $\mathcal{H}_1$) mais on sait dire qu'elle tend vers 1 quand n tend vers l'infini : on dit que le test est **consistant**.
Le principe habituel pour concevoir un test consiste à définir une variable auxiliaire T à valeurs réelles, fonction de l'échantillon, dont on connaît (ou contrôle) la loi sous $\mathcal{H}_0$ et qui a un comportement très différent sous $\mathcal{H}_1$. Typiquement, sous $\mathcal{H}_0$, T a une loi ν (ou converge en loi vers ν quand la taille n de l'échantillon tend vers l'infini) et, sous $\mathcal{H}_1$, T diverge vers $+\infty$ en probabilité avec n. On choisit alors un intervalle borné $A \subset \mathbb{R}$ (*région d'acceptation* ou, plus proprement, *région de non-rejet*) telle que $\nu(A) > 1-\alpha$, et le test est : « si $T \notin A$, on rejette $\mathcal{H}_0$ ». Sous $\mathcal{H}_0$, $\mathbb{P}(T \in A) = \nu(A) = 1-\alpha$ donc ce test est de niveau α. Et, sous $\mathcal{H}_1$, $\mathbb{P}(T \in A) \to_n 0$ car A est bornée et $T \to \infty$ (en probabilité), donc ce test est consistant.
NB. On note que, bizarrement, un intervalle de confiance de niveau $1-\alpha$ fournit un test de niveau α. Le mot *niveau* a donc des significations bien différentes selon les cas. Pour un intervalle de confiance, c'est le *niveau de confiance* ; pour un test, c'est le *niveau de risque* (de première espèce).

Remarque générale. Dans la suite, les énoncés sont souvent exacts dans le cas gaussien et approximatifs (pour n grand) dans un cadre très général. Il convient en pratique de veiller à ce que l'approximation gaussienne soit justifiée. Les énoncés pratiques sont donc assortis de limites d'application (valeur de n assez grande, etc.) que je ne précise pas toujours ci-après, d'autant qu'elles devraient théoriquement dépendre de la loi réellement suivie par les variables. Pour donner un ordre d'idée, on peut signaler qu'avant l'essor de l'informatique, une méthode populaire pour simuler des variables aléatoires de loi $\mathcal{N}(0,1)$, appelée « sum-of-twelve method », consistait à calculer $U_1 + \cdots + U_{12} - 6$ où $U_1, \ldots, U_{12}$ sont i.i.d. de loi uniforme sur $[0,1]$ (donc de moyenne $\frac{1}{2}$ et de variance $\frac{1}{12}$). L'approximation était déjà satisfaisante pour nombre d'usages (et d'un coût algorithmique très modéré).

2 Moyenne et variance

2.1 Estimation

On estime en général la moyenne d'une loi à l'aide de la moyenne empirique

$$\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

C'est un estimateur sans biais, et par la loi des grands nombres il est fortement consistant.
Si la moyenne m est connue, on peut estimer la variance à l'aide de

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - m)^2,$$

qui est sans biais et fortement consistant. En général, m est inconnue et on estime alors la variance à l'aide de la variance empirique

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}\big(X_k - \overline{X}_n\big)^2.$$

La normalisation par $n-1$ le rend sans biais (calcul laissé en exercice). De plus (en développant), cet estimateur est fortement consistant. Normaliser par n donne un autre estimateur fortement consistant qui peut aussi être utilisé.
Ces formules s'adaptent pour des variables à valeurs dans $\mathbb{R}^d$: la moyenne empirique se définit de même, et la matrice de covariance empirique est

$$\hat{\Gamma}_n = \frac{1}{n-1}\sum_{k=1}^{n}\big(X_k - \overline{X}_n\big)^t\big(X_k - \overline{X}_n\big).$$

2.2 Régions de confiance

Variance connue. Si $\mu = \mathcal{N}(m, \sigma^2)$, alors $\overline{X}_n \sim \mathcal{N}(m, \sigma^2/n)$, donc $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m) \sim \mathcal{N}(0,1)$. On a donc

$$\mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\big|\overline{X}_n - m\big| < A\right) = 1 - \alpha$$

où A est tel que

$$\int_A^\infty e^{-t^2/2}\frac{dt}{\sqrt{2\pi}} = \frac{\alpha}{2}$$

(si $\alpha = 0.05$, $A \simeq 1.96$) ce qui donne

$$\mathbb{P}\left(m \in \left[\overline{X}_n - \frac{\sigma A}{\sqrt{n}}, \overline{X}_n + \frac{\sigma A}{\sqrt{n}}\right]\right) = 1 - \alpha.$$

Autrement dit, $[\overline{X}_n - \frac{\sigma A}{\sqrt{n}}, \overline{X}_n + \frac{\sigma A}{\sqrt{n}}]$ est un intervalle de confiance de niveau $1-\alpha$ pour m. On a choisi cet intervalle symétrique du fait de la symétrie de la loi μ.
Par le théorème central limite, c'est de plus un intervalle de confiance asymptotique pour m dès que μ a pour variance σ^2.
Dans $\mathbb{R}^d$, si $\mu = \mathcal{N}(m, \sigma^2 I_d)$ alors $\overline{X}_n \sim \mathcal{N}(m, \sigma^2/nI_d)$, et comme cette loi est invariante par les rotations de centre m il est naturel de chercher une région de confiance qui soit une boule. On a

$$\frac{1}{\sigma^2}\|\overline{X}_n - m\|^2 = \frac{(X_1 - m)^2}{\sigma^2} + \cdots + \frac{(X_n - m)^2}{\sigma^2} \sim \chi^2_n$$

donc si $R_n > 0$ est telle que $\mathbb{P}(Z_n < R_n) = 1 - \alpha$ où $Z_n \sim \chi^2_n$ (R_n est de l'ordre de $1/n$), alors

$$\mathbb{P}(\|\overline{X}_n - m\| \le \sigma\sqrt{R_n}) = 1 - \alpha,$$

de sorte que la boule de centre $\overline{X}_n$ et de rayon $\sigma\sqrt{R_n}$ est une région de confiance de niveau $1-\alpha$ pour m.
Si maintenant $\mu = \mathcal{N}(m, \Gamma)$, et que A est une matrice carrée telle que $\Gamma = A^tA$, alors $A^{-1}(\overline{X}_n - m) \sim \mathcal{N}(0, 1/nI_d)$ donc (avec la même valeur R_n que ci-dessus)

$$\mathbb{P}(\|A^{-1}(\overline{X}_n - m)\| \le \sqrt{R_n}) = 1 - \alpha$$

d'où

$$\mathbb{P}(m \in A \cdot B(\overline{X}_n, \sqrt{R_n})) = 1 - \alpha.$$

Autrement dit, l'ellipsoïde image de la boule de centre $\overline{X}_n$ et de rayon $\sqrt{R_n}$ par l'application A est une région de confiance pour m de niveau $1-\alpha$.

Moyenne inconnue. Donnons un intervalle de confiance pour la variance lorsque la moyenne est inconnue. On utilise la conséquence importante du théorème de Cochran qui suit :

Proposition 3. Si $X_1, \dots, X_n$ des variables aléatoires i.i.d. de loi $\mathcal{N}(m, \sigma^2)$, alors les variables aléatoires $\overline{X}_n$ et S_n^2 sont indépendantes,

$$\overline{X}_n \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right) \quad \text{et} \quad \frac{n-1}{\sigma^2} S_n^2 \sim \chi^2_{n-1}.$$

Démonstration. On considère la décomposition $\mathbb{R}^d = E \oplus E^\perp$ où $E = \mathbb{R}\mathbf{1}$ ($\mathbf{1}$ étant le vecteur de $\mathbb{R}^n$ dont toutes les composantes valent 1). En notant $X = {}^t\begin{pmatrix} X_1 & \cdots & X_n \end{pmatrix}$, on a :

$$P_E(X) = \frac{X \cdot \mathbf{1}}{\|\mathbf{1}\|^2}\mathbf{1} = \begin{pmatrix} \overline{X}_n \\ \vdots \\ \overline{X}_n \end{pmatrix} \quad \text{et} \quad P_{E^\perp}(X) = X - P_E(X) = \begin{pmatrix} X_1 - \overline{X}_n \\ \vdots \\ X_n - \overline{X}_n \end{pmatrix},$$

donc $\overline{X}_n$ ne dépend que de $P_E(X)$ et $S_n^2 = \frac{1}{n-1}\|P_{E^\perp}(X)\|^2$ ne dépend que de $P_{E^\perp}(X)$. Le théorème de Cochran conclut, avec le fait que $E^\perp$ est de dimension $n-1$. □

Donc si $\mu = \mathcal{N}(m, \sigma^2)$, en choisissant $0 \leq a_n < b_n$ tels que $\mathbb{P}(a_n < Z_n < b_n) = 1 - \alpha$ où $Z_n \sim \chi^2_{n-1}$, on a

$$\mathbb{P}\left(\frac{n-1}{\sigma^2} S^2_{n-1} \in [a, b]\right) = 1 - \alpha$$

et donc

$$\mathbb{P}\left(\sigma^2 \in \left[\frac{n-1}{b} S_n^2, \frac{n-1}{a} S_n^2\right]\right) = 1 - \alpha,$$

ce qui donne un intervalle de confiance pour σ (sans connaître m).

> **Remarque**. La loi χ^2_n n'est pas symétrique (elle est portée par $\mathbb{R}_+$). Un choix naturel d'intervalle de confiance consiste à prendre a_n, b_n tels que $\mathbb{P}(Z_n < a_n) = \frac{\alpha}{2}$ et $\mathbb{P}(Z_n > b_n) = \frac{\alpha}{2}$. Notons que a_n et b_n tendent vers $+\infty$ avec n.

Variance inconnue. On suppose $\mu = \mathcal{N}(m, \sigma^2)$ avec m et σ^2 inconnues. L'intervalle de confiance pour m donné précédemment fait intervenir σ. Pour éviter cela, on peut borner σ, ou l'approcher.
Si $\mu = \mathcal{B}(p)$ (réponses oui/non à un sondage), alors $\sigma^2 = p(1-p) \leq \frac{1}{4}$. Reprenant l'intervalle de confiance asymptotique de niveau $1-\alpha$ précédent, on en déduit l'intervalle asymptotique suivant (plus large, mais de peu pour p proche de $1/2$) :

$$\left[\overline{X}_n - \frac{A}{2\sqrt{n}}, \overline{X}_n + \frac{A}{2\sqrt{n}}\right].$$

Si μ est quelconque de carré intégrable, on souhaite remplacer σ^2 par S_n^2 dans l'intervalle de confiance asymptotique donné plus haut. Pour cela, il suffit de justifier qu'on peut le faire dans le théorème central limite :

$$\frac{\sqrt{n}}{\sqrt{S_n^2}}(\overline{X}_n - m) \xrightarrow[n]{\text{(loi)}} \mathcal{N}(0,1).$$

C'est effectivement vrai car $(S_n^2)_n$ converge presque-sûrement vers σ^2 et on dispose du

Proposition 4 – *Lemme de Slutzky.* Si les suites $(X_n)_n$ et $(Y_n)_n$ convergent respectivement en loi vers une variable aléatoire X et vers une constante c, alors la suite $(X_n, Y_n)_n$ des couples converge en loi vers (X, c). Ainsi, $f(X_n, Y_n)$ converge en loi vers $f(X, c)$ quand $n \to \infty$ pour toute fonction continue f.

Ainsi, l'intervalle

$$\left[\overline{X}_n - \frac{A\sqrt{S_n^2}}{\sqrt{n}}, \overline{X}_n + \frac{A\sqrt{S_n^2}}{\sqrt{n}}\right],$$

où A est tel que $\mathbb{P}(|Z| \leq A) = 1 - \alpha$ pour $Z \sim \mathcal{N}(0,1)$, est un intervalle de confiance asymptotique de niveau $1-\alpha$ pour m, qui ne dépend pas de σ.

Si $\mu = \mathcal{N}(m, \sigma^2)$, cet intervalle de confiance n'est pas exact puisque la loi de $\frac{\sqrt{n}}{\sqrt{S_n^2}}(\overline{X}_n - m)$ n'est pas gaussienne. L'importance pratique de cette situation justifie l'intérêt de chercher plutôt un intervalle de confiance exact. Par la Proposition 3, $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m) \sim \mathcal{N}(0,1)$, $\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$ et ces variables aléatoires sont indépendantes, donc (par définition)

$$\frac{\sqrt{n}}{\sqrt{S_n^2}}(\overline{X}_n - m) = \frac{\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m)}{\sqrt{\frac{1}{\sigma^2}S_n^2}} \sim t_{n-1}$$

(loi de Student à $n-1$ degrés de liberté, voir appendice). Si on choisit A_n tel que $\mathbb{P}(|Z_n| \leq A_n) = 1-\alpha$ où $Z_n \sim t_{n-1}$, alors l'intervalle donné plus haut, avec A_n au lieu de A, est un intervalle de confiance (exact) de niveau $1-\alpha$ pour m. Quand $n \to \infty$, $A_n \to A$, donc cet intervalle de confiance exact (mais plus difficile à calculer) n'est intéressant que pour de petites valeurs de n.

2.3 Tests de valeur

Moyenne Pour une valeur m_0 donnée, on souhaite tester l'hypothèse $\mathcal{H}_0$: « $m = m_0$ » contre l'une des hypothèses $\mathcal{H}_1$: « $m \neq m_0$ », $\mathcal{H}_1^>$: « $m > m_0$ » ou $\mathcal{H}_1^<$: « $m < m_0$ ».
Supposons σ^2 inconnue (le cas où σ^2 est connue fonctionne de même, avec $\mathcal{N}(0,1)$ au lieu de t_{n-1} et σ^2 au lieu de S_n^2). On a vu que, sous $\mathcal{H}_0$,

$$T := \frac{\sqrt{n}}{\sqrt{S_n^2}}(\overline{X}_n - m_0) = \frac{\frac{\sqrt{n}}{\sigma}(\overline{X}_n - m_0)}{\sqrt{\frac{1}{\sigma^2}S_n^2}} \sim t_{n-1}$$

donc, sous $\mathcal{H}_0$, $\mathbb{P}(|T| \leq A_n) = 1-\alpha$ où A_n est tel que $\mathbb{P}(Z_n \leq A_n) = 1-\alpha$ si $Z_n \sim t_{n-1}$.
Le test « On rejette $\mathcal{H}_0$ si $|T| > A_n$ » est un test de niveau α de $\mathcal{H}_0$ contre l'une quelconque des hypothèses $\mathcal{H}_1$, $\mathcal{H}_1^>$ ou $\mathcal{H}_1^<$.
De plus, sous $\mathcal{H}_1^>$, $\overline{X}_n - m_0 \to m - m_0 > 0$ donc $T \to +\infty$ p.s., ce qui donne $\mathbb{P}(|T| \leq A_n) \to_n 0$. Sous $\mathcal{H}_1^<$, $T \to -\infty$ p.s., et sous $\mathcal{H}_1$ l'un ou l'autre a lieu. Dans tous les cas, le test est consistant.
La différence entre les hypothèses alternatives est dans la puissance du test. Si l'hypothèse alternative est vraie, un bon test doit rejeter l'hypothèse avec grande probabilité. On sait que c'est vrai asymptotiquement. Préciser l'hypothèse alternative permet de raffiner légèrement le test :

- Sous $\mathcal{H}_1^>$, l'hypothèse est rejetée du fait de grandes valeurs de T, donc il est inutile d'avoir une région de rejet située à gauche de m. Le test sera meilleur si on choisit la zone d'acceptation de la forme $]-\infty, A]$: « On rejette $\mathcal{H}_0$ si $T > A_n^+$ » où A_n^+ est tel que $\mathbb{P}(Z_n \leq A_n^+) = 1-\alpha$ si $Z \sim t_{n-1}$. On parle de test *unilatéral*.
- Si l'hypothèse alternative est $\mathcal{H}_1^<$, on choisit de même le test « On rejette $\mathcal{H}_0$ si $T > A_n^-$ » où A_n^- est tel que $\mathbb{P}(Z_n \geq A_n^-) = 1-\alpha$ si $Z \sim t_{n-1}$. (En fait, $A_n^- = -A_n^+$)
- Contre $\mathcal{H}_1$, on utilise la version initiale : un test bilatéral.

Variance On souhaite tester $\mathcal{H}_0$: « $\sigma = \sigma_0$ » contre $\mathcal{H}_1$: « $\sigma \neq \sigma_0$ », $\mathcal{H}_1^<$: « $\sigma < \sigma_0$ » ou $\mathcal{H}_1^>$: « $\sigma > \sigma_0$ ».
On applique la même principe que ci-dessus avec la variable

$$C = \frac{n-1}{\sigma_0^2}S_n^2.$$

On sait que, sous $\mathcal{H}_0$, $C \sim \chi^2_{n-1}$, et C sera plus grande sous $\mathcal{H}_1^>$, donc :

- pour un test bilatéral ($\mathcal{H}_0$ contre $\mathcal{H}_1$), on choisit a_n, b_n tels que $\mathbb{P}(a_n < Z_n < b_n) = 1-\alpha$ si $Z_n \sim \chi^2_{n-1}$;
- pour un test de $\mathcal{H}_0$ contre $\mathcal{H}_1^>$, on choisit a_n tel que $\mathbb{P}(a_n < Z_n) = 1-\alpha$;
- pour un test de $\mathcal{H}_0$ contre $\mathcal{H}_1^<$, on choisit b_n tel que $\mathbb{P}(Z_n > b_n) = 1-\alpha$.

Alors le test « On rejette $\mathcal{H}_0$ si $C \notin [a_n, b_n]$ » (resp. $C < a_n$, $C > b_n$) est un test de niveau α de $\mathcal{H}_0$ contre $\mathcal{H}_1$ (resp. $\mathcal{H}_1^>$, $\mathcal{H}_1^<$).
De plus, on peut justifier que ce test est consistant. De même qu'avant, le premier test est en fait consistant pour chacune des trois hypothèses, mais moins puissant que les versions unilatérales dans le cas de $\mathcal{H}_1^>$ et $\mathcal{H}_1^<$. On note qu'ici $C \sim n\frac{\sigma^2}{\sigma_0^2}$ tend vers $+\infty$ p.s. quel que soit σ, donc la consistance n'est pas aussi directe que pour le test de valeur de la moyenne. Elle se déduit du fait que $a_n \sim b_n \sim n$ et $\frac{C}{n} \to_n \frac{\sigma^2}{\sigma_0^2}$ p.s., donc sous $\mathcal{H}_1$ l'intervalle $[a_n/n, b_n/n]$ ne contient presque sûrement plus $\frac{C}{n}$ pour n grand.

2.4 Tests de comparaison

On suppose que l'on dispose de deux échantillons $X_1, \dots, X_n$ de loi $\mathcal{N}(m_x, \sigma_X^2)$ et $Y_1, \dots, Y_p$ de loi $\mathcal{N}(m_Y, \sigma_Y^2)$ indépendants entre eux. On souhaite tester l'égalité de leurs moyennes et variances.

Comparaison de moyennes Pour le test qui suit, **on suppose les variances égales** (on verra juste après comment tester cette hypothèse) : $\sigma = \sigma_X = \sigma_Y$. En revanche, on suppose σ inconnue.
On souhaite tester $\mathcal{H}_0$: « $m_X = m_Y$ » contre $\mathcal{H}_1$: « $m_X \neq m_Y$ », ou contre $\mathcal{H}_1^{X>Y}$: « $m_X > m_Y$ », ou contre $\mathcal{H}_1^{X<Y}$: « $m_X < m_Y$ ».
Comme $m_X - m_Y$ est estimé par $\overline{X}_n - \overline{Y}_p$, on peut chercher un test basé sur les valeurs de cette variable. Elle suit la loi $\mathcal{N}(m_X - m_Y, \frac{\sigma^2}{n} + \frac{\sigma^2}{p})$. Il faut se ramener à une loi connue, ne dépendant pas de σ. Si σ était connue, on dirait que, sous $\mathcal{H}_0$,

$$\frac{\sqrt{\frac{1}{n} + \frac{1}{p}}}{\sigma}\big(\overline{X}_n - \overline{Y}_p\big) \sim \mathcal{N}(0,1).$$

Ici, on va utiliser comme d'habitude S_X^2 et S_Y^2 (les variances empiriques) pour remplacer σ.
On applique le théorème de Cochran à la décomposition orthogonale $\mathbb{R}^{n+p} = E_X \oplus E_Y \oplus F$ et au vecteur Z, où

$$E_X = \mathbb{R}\begin{pmatrix}1\\ \vdots\\ 1\\ 0\\ \vdots\\ 0\end{pmatrix}, \quad E_Y = \mathbb{R}\begin{pmatrix}0\\ \vdots\\ 0\\ 1\\ \vdots\\ 1\end{pmatrix}, \quad F = (E_X + E_Y)^\perp \quad \text{et } Z = \begin{pmatrix}X_1\\ \vdots\\ X_n\\ Y_1\\ \vdots\\ Y_p\end{pmatrix}.$$

Alors (comme dans la proposition 3) $\overline{X}_n$ ne dépend que de $P_{E_X}(Z)$, $\overline{Y}_p$ ne dépend que de $P_{E_Y}(Z)$ et $\|P_F(Z)\|^2 = (n-1)S_X^2 + (p-1)S_Y^2$, donc ces trois quantités sont indépendantes, et la dernière a la loi de $\sigma^2 C$ où $C \sim \chi^2_{n+p-2}$ car $\dim F = n + p - 2$. Par suite, la variable aléatoire

$$T = \frac{\sqrt{\frac{1}{n} + \frac{1}{p}}\big(\overline{X}_n - \overline{Y}_p\big)}{\sqrt{\frac{(n-1)S_X^2 + (p-1)S_Y^2}{n+p-2}}}$$

suit, sous $\mathcal{H}_0$, la loi t_{n+p-2}. On en déduit les tests de niveau α :
– De $\mathcal{H}_0$ contre $\mathcal{H}_1$: « On rejette l'hypothèse $m_X = m_Y$ si $|T| > A$ », où A est tel que $\mathbb{P}(C > A) = \alpha/2$ si $C \sim \chi^2_{n+p-2}$.
– De $\mathcal{H}_0$ contre $\mathcal{H}_1^{X>Y}$: « On rejette l'hypothèse $m_X = m_Y$ si $T > A$ », où A est tel que $\mathbb{P}(C > A) = \alpha$.
– De $\mathcal{H}_0$ contre $\mathcal{H}_1^{X<Y}$: « On rejette l'hypothèse $m_X = m_Y$ si $T < -A$ », où A est tel que $\mathbb{P}(C > A) = \alpha$.
Si $m_X \neq m_Y$, T diverge vers $\pm\infty$ p.s. quand $n, p \to \infty$ d'où la consistance (comme pour les tests précédents).

Comparaison de variances On souhaite tester l'hypothèse $\mathcal{H}_0$: « $\sigma_X = \sigma_Y$ » contre $\mathcal{H}_1$: « $\sigma_X \neq \sigma_Y$ (ou … ou …).
On se rappelle (proposition 3) que $\frac{n-1}{\sigma_X^2} S_X^2 \sim \chi^2_{n-1}$, $\frac{p-1}{\sigma_Y^2} S_Y^2 \sim \chi^2_{p-1}$ et ces variables sont indépendantes par hypothèse, donc sous $\mathcal{H}_0$ (c-à-d. si $\sigma_X = \sigma_Y$), la variable

$$F = \frac{\frac{n-1}{\sigma_X^2} S_X^2/(n-1)}{\frac{p-1}{\sigma_Y^2} S_Y^2/(p-1)} = \frac{\sigma_Y^2}{\sigma_X^2}\frac{S_X^2}{S_Y^2} = \frac{S_X^2}{S_Y^2}$$

suit (par définition) la loi $F(n-1, p-1)$ (loi de Fischer-Snedecor).
On a alors le test de niveau α suivant : « On rejette l'hypothèse $\sigma_X = \sigma_Y$ si $F \notin [a, b]$ » où a, b sont tels que $\mathbb{P}(Z < a) = \mathbb{P}(Z > b) = \frac{\alpha}{2}$ si $Z \sim F(n-1, p-1)$. (Et on a les tests unilatéraux pour les hypothèses alternatives correspondantes).

3 Lois discrètes : tests du χ^2

3.1 Test du χ^2 d'adéquation

On suppose que μ est une loi sur l'ensemble fini $\{1,\dots,r\}$. Autrement dit, il existe un vecteur $(p_1,\dots,p_r)$ de probabilité ($p_k \geq 0$ et $p_1+\cdots+p_r=1$) tel que

$$\mathbb{P}(X_1=k)=p_k \quad \text{pour } k=1,\dots,r.$$

Notons que l'on peut estimer $p_1,\dots,p_r$ à l'aide de

$$\hat{p}_k=\frac{N_k}{n} \quad \text{où } N_k=\mathrm{Card}\{1\leq i\leq n | X_i=k\} \quad \text{pour } k=1,\dots,r.$$

(N_k est l'*effectif* de la *classe* k). Ceci revient à estimer la moyenne des vecteurs aléatoires

$$Z_i=\begin{pmatrix}\mathbf{1}_{\{X_i=1\}}\\ \vdots \\ \mathbf{1}_{\{X_i=k\}}\end{pmatrix}.$$

On souhaite tester l'hypothèse $\mathcal{H}_0$: « $p=q$ » où $q=(q_1,\dots,q_r)$ est un vecteur de probabilité donné, contre $\mathcal{H}_1$: « $p\neq q$ ».
On pourrait penser à un test portant sur la valeur de $\sum_{k=1}^r(\hat{p}_k-q_k)^2$. Afin de définir une statistique qui ne dépende pas de q (et dont on peut donc faire des tables pour ajuster le niveau du test), il faut cependant pondérer les termes d'une certaine manière : le test portera sur

$$D(\hat{p},q)=\sum_{k=1}^r\frac{(\hat{p}_k-q_k)^2}{q_k}=\frac{1}{n}\sum_{k=1}^r\frac{(N_k-nq_k)^2}{nq_k}.$$

On prouve à l'aide du théorème de Cochran (cf. ci-dessous) que :

$$\text{sous } \mathcal{H}_0, \quad nD(\hat{p},q)\xrightarrow[n]{\text{(loi)}}\chi^2_{r-1}.$$

On remarque aussi que, si $q\neq p$, $D(\hat{p},q)$ admet une limite >0 p.s. donc $nD(\hat{p},q)\to_n+\infty$ p.s.
D'où un test consistant de niveau α : « On rejette l'hypothèse $p=q$ si $nD(\hat{p},q)>A$ » où A est tel que $\mathbb{P}(C>A)=\alpha$ avec $C\sim\chi^2_{r-1}$.
En pratique, on demande souvent d'avoir $N_k\geq 5$ pour tout k afin que le test soit significatif. Si ce n'est pas le cas, on peut par exemple regrouper des classes.

Remarque. Ce test porte sur l'adéquation avec une loi à support fini : par exemple, pour tester si $X_1,\dots,X_n$ suivent une loi binomiale de paramètre (n,p) donné. Mais on peut aussi l'utiliser pour tester l'adéquation avec une loi à support infini en regroupant les valeurs pour se ramener à un support fini : par exemple, pour une loi sur $\mathbb{R}$, les classes pourraient être $]-\infty,-N]$, $]-N,-N+1[,\dots,]N-1,N]$, $]N,+\infty[$, et pour une loi sur $\mathbb{N}$, $\{0\},\dots,\{N\}$, $\{N+1,\dots\}$ (avec toujours la possibilité de regrouper des classes afin d'éviter les effectifs trop faibles).

Preuve du test. On se place désormais sous $\mathcal{H}_0$: $p=q$. On introduit les vecteurs

$$Y_i=\begin{pmatrix}\frac{\mathbf{1}_{\{X_i=1\}}-q_1}{\sqrt{q_1}}\\ \vdots \\ \frac{\mathbf{1}_{\{X_i=r\}}-q_r}{\sqrt{q_r}}\end{pmatrix}.$$

Ces variables aléatoires $Y_1,\dots,Y_n$ sont i.i.d. Sous $\mathcal{H}_0$, elles sont centrées (espérance nulle) et le théorème central limite (multidimensionnel) donne

$$\frac{1}{\sqrt{n}}\big(Y_1+\cdots+Y_n\big)\xrightarrow[n]{\text{(loi)}}W$$

où $W\sim\mathcal{N}(0,\Gamma)$, Γ étant la matrice de covariance de Y_1. Comme $x\mapsto\|x\|^2$ est continue sur $\mathbb{R}^r$, on en déduit

$$nD(\hat{p},q)=\left\|\frac{1}{\sqrt{n}}\big(Y_1+\cdots+Y_n\big)\right\|^2\xrightarrow[n]{\text{(loi)}}\|W\|^2.$$

Il reste à déterminer la loi de $\|W\|^2$ si $W \sim \mathcal{N}(0,\Gamma)$. Calculons $\Gamma = \mathrm{Var}(Y_1)$. Pour $k = 1,\dots,r$,

$$\mathrm{Var}\left(\frac{\mathbf{1}_{\{X_1=k\}} - q_k}{\sqrt{q_k}}\right) = \frac{1}{q_k}\mathrm{Var}\left(\mathbf{1}_{\{X_1=k\}}\right) = \frac{q_k(1-q_k)}{q_k} = 1 - q_k$$

et, pour $1 \le k \ne l \le n$,

$$\mathrm{Cov}\left(\frac{\mathbf{1}_{\{X_1=k\}} - q_k}{\sqrt{q_k}}, \frac{\mathbf{1}_{\{X_1=l\}} - q_l}{\sqrt{q_l}}\right) = \frac{1}{\sqrt{q_k q_l}}\mathrm{Cov}\left(\mathbf{1}_{\{X_1=k\}}, \mathbf{1}_{\{X_1=l\}}\right) = \frac{-q_k q_l}{\sqrt{q_k q_l}} = -\sqrt{q_k q_l}$$

donc

$$\Gamma = I - \sqrt{q}^t(\sqrt{q}) \quad \text{où} \quad \sqrt{q} = \begin{pmatrix}\sqrt{q_1}\\ \vdots \\ \sqrt{q_r}\end{pmatrix}.$$

On note que $\sqrt{q}$ est un vecteur unitaire. Si bien que Γ est la matrice de la projection orthogonale sur $\sqrt{q}^\perp$: pour tout $x \in \mathbb{R}^r$, $P_{\sqrt{q}^\perp}(x) = x - (x\cdot\sqrt{q})\sqrt{q} = x - \sqrt{q}^t(\sqrt{q})x = \Gamma x$. En particulier, $\Gamma = \Gamma^2 = \Gamma^t\Gamma$ donc W a même loi que $\Gamma U = P_{\sqrt{q}^\perp}(U)$ où $U \sim \mathcal{N}(0, I_r)$. Et, par le théorème de Cochran, $\|W\|^2 \sim \chi^2_{r-1}$ car $\dim(\sqrt{p}^\perp) = r-1$. D'où la conclusion : la limite vue plus haut s'écrit

$$\text{sous } \mathcal{H}_0, \quad nD(\hat{p}, q) \xrightarrow[n]{\text{(loi)}} \chi^2_{r-1}.$$

3.2 Test du χ^2 d'indépendance

On suppose que μ est une loi sur $\{1,\dots,r\}\times\{1,\dots,s\}$, c'est-à-dire que l'échantillon est $(X_1,Y_1),\dots,(X_n,Y_n)$ où X_i est à valeurs dans $\{1,\dots,r\}$ et Y_i dans $\{1,\dots,s\}$.
On souhaite tester l'hypothèse $\mathcal{H}_0$: « X_1 et Y_1 sont indépendants » (autrement dit, μ est une loi produit) contre $\mathcal{H}_1$: « X_1 et Y_1 ne sont pas indépendants ».
S'il y a indépendance, alors $p_{kl} = \mathbb{P}(X_i = k, Y_i = l) = p_{k\cdot}p_{\cdot l}$ où $p_{k\cdot} = \mathbb{P}(X_i = k)$ et $p_{\cdot l} = \mathbb{P}(X_i = l)$. Or on peut estimer d'un côté p_{kl} par $\hat{p}_{kl} = \frac{N_{kl}}{n}$ et d'autre part $p_{k\cdot}$ et $p_{\cdot k}$ par $\hat{p}_{k\cdot} = \frac{N_{k\cdot}}{n}$ et $\hat{p}_{\cdot l} = \frac{N_{\cdot l}}{n}$ où

$$N_{kl} = \mathrm{Card}\{i | (X_i,Y_i) = (k,l)\}, \quad N_{k\cdot} = \mathrm{Card}\{i | X_i = k\} \quad \text{et} \quad N_{\cdot l} = \mathrm{Card}\{i | Y_i = l\}.$$

Le test utilise la quantité

$$D = \sum_{1\le k\le r}\sum_{1\le l\le s}\frac{(\hat{p}_{kl} - \hat{p}_{k\cdot}\hat{q}_{l\cdot})^2}{\hat{p}_{k\cdot}\hat{p}_{\cdot l}} = \frac{1}{n}\sum_{k,l}\frac{\left(N_{kl} - \frac{1}{n}N_{k\cdot}N_{\cdot l}\right)^2}{\frac{N_{k\cdot}N_{\cdot l}}{n}}$$

Le résultat (dont la preuve est délicate) est le suivant :

$$\text{sous } \mathcal{H}_0, \quad nD \xrightarrow[n]{\text{(loi)}} \chi^2_{(r-1)(s-1)}$$

et $nD \to +\infty$ p.s. sous $\mathcal{H}_1$. On a donc le test consistant de niveau α suivant : « On rejette l'hypothèse d'indépendance si $nD > A$ » où A est choisi tel que $\mathbb{P}(C > A) = \alpha$ si $C \sim \chi^2_{(r-1)(s-1)}$.
Là encore, on suppose en pratique que $N_{kl} \ge 5$ pour appliquer le test.

3.3 Autres tests du χ^2

Les tests d'adéquation et d'indépendance sont les plus utilisés. Il en existe beaucoup d'autres tests « du χ^2 » : test d'homogénéité (est-ce que les deux échantillons à valeurs dans un ensemble fini ont même loi ?), test de symétrie (est-ce que les couples (X_i,Y_i) et (Y_i,X_i) ont même loi (sur un ensemble fini) ?),...

Remarque. On parle de loi du chi-deux à r *degrés de libertés*. On peut comprendre cette expression : χ^2_r est la loi d'un vecteur formé de r composantes gaussiennes standard indépendantes (donc « libres »). Par exemple, $\frac{n-1}{\sigma}S_n^2 \sim \chi^2_{n-1}$ provient de la projection d'un vecteur gaussien sur un sous-espace de dimension $n-1$, ce qui se produit en estimant la moyenne par $\overline{X}_n$: les composantes $X_1 - \overline{X}_n, \dots, X_n - \overline{X}_n$ sont liées par la condition que leur somme est nulle. On peut retrouver, au moins intuitivement, les nombres de degrés de liberté dans les tests du χ^2 : pour l'adéquation, les composantes $p_1,\dots,p_r$ somment à 1, d'où la perte d'un degré de liberté ; et pour l'indépendance, la loi μ est donnée par celle de chaque marginale (sur $\{1,\dots,r\}$ et $\{1,\dots,s\}$), qui ne dépendent que de $r-1$ et $s-1$ paramètres (car elles somment à 1), d'où $(r-1)(s-1)$ vrais paramètres, ou degrés de liberté. On pourrait justifier que, pour effectuer un test d'adéquation avec une loi ayant plusieurs paramètres inconnus, chaque paramètre estimé réduit de 1 le nombre de degrés de liberté de la loi du χ^2 obtenue à la limite. On pourrait alors voir tous les tests du χ^2 comme des cas particulier du test d'adéquation, avec estimation de certains paramètres. La preuve (et l'énoncé) d'un résultat aussi général est malheureusement délicate.

4 Test de Kolmogorov-Smirnov

Le test du χ^2 d'adéquation permet la comparaison à une loi à support fini et éventuellement, on l'a signalé, à une loi plus générale en définissant des classes adaptées. Le test de Kolmogorov-Smirnov permet de tester l'adéquation à une loi continue sur $\mathbb{R}$, il est plus puissant qu'un tel test du χ^2 (qui ne va pas rejeter l'hypothèse si la loi réellement suivie donne la même probabilité aux classes que celle qui est testée).
On suppose que $X_1, \dots, X_n$ suivent une loi sur $\mathbb{R}$.
Le test utilise la **fonction de répartition empirique**

$$F_n : t \mapsto F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}} = \frac{\text{Card}\{1 \leq i \leq n | X_i \leq t\}}{n}.$$

(C'est la fonction de répartition de la loi empirique $\hat{\mu}$ définie précédemment)
Par la loi forte des grands nombres, pour tout t, $F_n(t)$ converge vers $F(t)$ p.s. On peut justifier que p.s., pour tout t, $F_n(t)$ converge vers $F(t)$ (noter la différence), en utilisant le fait que F_n et F sont continues à droite et croissantes. Et on peut aussi, en utilisant un théorème de Dini (pas le plus courant) et le fait que F_n et F ont des limites en $-\infty$ et $+\infty$, montrer que la convergence est uniforme : p.s., $\sup_{\mathbb{R}} |F_n - F| \to 0$.

Théorème 4. On suppose que μ est diffuse, c'est-à-dire sans atome (c'est notamment le cas si μ a une densité). Alors

$$\sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n]{\text{(loi)}} KS,$$

où KS est la loi de Kolmogorov-Smirnov : c'est une loi sur $\mathbb{R}_+$ qui ne dépend pas de μ.

Il existe des tables de la loi de Kolmogorov-Smirnov (dans la boîte à outils StixBox par exemple).
Soit G la fonction de répartition d'une loi diffuse ν sur $\mathbb{R}$. On souhaite tester l'hypothèse « $\mu = \nu$ » contre « $\mu \neq \nu$ ».
On peut donc faire le test suivant, consistant et de niveau α : « On rejette l'hypothèse $\mu = \nu$ si $\sqrt{n}\sup_{\mathbb{R}} |F_n - G| > A$ » où A est tel que $\mathbb{P}(Z > A) = \alpha$ si Z suit la loi de Kolmogorov-Smirnov.
Notons qu'en pratique le calcul de la distance $\sup_n |F_n - G|$ se résume au calcul du maximum des différences aux points $X_1, \dots, X_n$ (à gauche et à droite) du fait de la croissance des fonctions.

5 Régression linéaire

5.1 Cadre général

On suppose que l'on dispose de données $x_1, \dots, x_n \in \mathbb{R}^p$ et $y_1, \dots, y_n \in \mathbb{R}$ liées par une relation de la forme

$$y_k = f(x_k) + \varepsilon_k,$$

où f est une fonction affine $\mathbb{R}^p \to \mathbb{R}$ inconnue et $\varepsilon_1, \dots, \varepsilon_n$ sont des erreurs de mesures, inconnues elles aussi. L'objectif est de déterminer la fonction f à partir des données.
Sous forme matricielle, ceci s'écrit

$$Y = X\Theta + \varepsilon,$$

où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ & & \cdots & \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \quad \Theta = \begin{pmatrix} b \\ a_1 \\ \vdots \\ a_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Exemple. On suppose que deux grandeurs mesurables expérimentalement, U et V, réelles, sont reliées par une relation de la forme $V = CU^\alpha e^{-\beta U}$ et l'on cherche à évaluer les paramètres C, α, β. Comme on a $\ln V = \ln C + \alpha \ln U - \beta U$, le modèle se prête à une régression linéaire avec $p = 2$: si les observations sont $V_1, \dots, V_n$ et $U_1, \dots, U_n$, alors avec la notation précédente

$$Y = \begin{pmatrix} \ln V_1 \\ \vdots \\ \ln V_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & \ln U_1 & U_1 \\ & \cdots & \\ 1 & \ln U_n & U_n \end{pmatrix}, \quad \Theta = \begin{pmatrix} b \\ a_1 \\ a_2 \end{pmatrix},$$

et si $Y = X\Theta$ alors finalement $\ln C = b$, $\alpha = a_1$ et $-\beta = a_2$. Le terme ε tient compte des erreurs de mesure.

On suppose le modèle identifiable, c'est-à-dire qu'il existe un seul vecteur Θ solution. Cela revient à dire que X est injective, et donc que X est de rang maximal, c'est-à-dire $p+1$: les colonnes de X sont indépendantes. Un modèle probabiliste consiste à supposer que $\varepsilon_1, \dots, \varepsilon_n$ sont des variables aléatoires indépendantes, centrées et de variances $\sigma_1^2, \dots, \sigma_n^2$ finies, et $y_1, \dots, y_n$ sont alors des variables aléatoires. La proposition suivante est essentiellement algébrique et ne dépend pas du modèle choisi.

Proposition 5. Il existe un unique vecteur aléatoire $\widehat{\Theta}$ fonction de $y_1, \dots, y_n$ qui minimise le risque quadratique $R : \Theta \mapsto \mathbb{E}\left[\|Y - X\widehat{\Theta}\|^2\right]$: c'est le vecteur tel que $X\widehat{\Theta}$ est la projection orthogonale de Y sur $E = \mathrm{Im}(X)$. On l'appelle *estimateur des moindres carrés* et il est donné par

$$\widehat{\Theta} = ({}^tXX)^{-1}({}^tX)Y.$$

Démonstration. Notons déjà que ${}^tXX$ a même noyau que X (si ${}^tXX\Theta = 0$, alors multiplier par ${}^t\Theta$ à gauche donne $\|X\Theta\|^2 = 0$ donc $X\Theta = 0$), donc est injective, et donc inversible, étant carrée.
La première partie est en fait évidente car vraie presque-sûrement : $\inf_\Theta \|Y - X\Theta\|$ est la distance de Y à E, réalisée uniquement en $\widehat{\Theta}$.
Et la deuxième partie s'obtient ainsi : on a $Y = X\widehat{\Theta} + Z$ où $Z \in E^\perp$, d'où ${}^tXY = {}^tXX\widehat{\Theta} + {}^tXZ = ({}^tXX)\widehat{\Theta}$ et ${}^tXX$ est inversible. □

Cas particulier. Le cas où les x_k sont réels ($p = 1$) admet une expression simple : alors $\widehat{\Theta} = \begin{pmatrix} \widehat{b} \\ \widehat{a} \end{pmatrix}$ avec

$$\widehat{a} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$$

où $\overline{z} = \frac{1}{n}\sum_{k=1}^n z_k$ (avec $z_k = x_k y_k$, x_k, etc.), et $\widehat{b}$ se déduit de $\overline{y} = \widehat{a}\overline{x} + \widehat{b}$.

Ce vecteur $\widehat{\Theta}$ est un estimateur de Θ, à l'aide duquel on définit

$$\widehat{Y} = X\widehat{\Theta},$$

estimateur de $X\Theta$, et

$$\widehat{\varepsilon} = Y - \widehat{Y},$$

estimateur de ε.
Préciser le modèle permet d'obtenir des régions de confiance et de réaliser des tests.

5.2 Cas gaussien

On suppose maintenant que $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$: les erreurs sont i.i.d. gaussiennes centrées et de même variance, en général inconnue. Alors $Y = X\Theta + \varepsilon \sim \mathcal{N}(X\Theta, \sigma^2 I_n)$.

Remarque. On note que Y a pour densité

$$y \mapsto f_\Theta(y) = \frac{1}{(2\pi\sigma^2)^{r/2}} e^{-\|y - X\Theta\|^2/(2\sigma^2)}$$

et donc $f_\Theta(Y)$ est maximum quand $\|Y - X\Theta\|$ est minimum, c'est-à-dire pour $\Theta = \widehat{\Theta}$: autrement dit, l'estimateur des moindres carrés $\widehat{\Theta}$ est un estimateur du maximum de vraisemblance.

Comme $\widehat{\Theta} = ({}^tXX)^{-1}({}^tX)Y$, et $\left(({}^tXX)^{-1}({}^tX)\right){}^t\left(({}^tXX)^{-1}({}^tX)\right) = ({}^tXX)^{-1}$, on a

$$\widehat{\Theta} \sim \mathcal{N}\left(\Theta, \sigma^2({}^tXX)^{-1}\right).$$

On pourrait donc donner un ellipsoïde de confiance pour $\widehat{\Theta}$ (voir Section ? ?) ou des intervalles de confiance pour ses composantes, mais qui dépendent de σ. Pour l'éviter, on estime σ^2 à l'aide de $\widehat{\varepsilon} = Y - \widehat{Y}$.

Proposition 6. Les estimateurs $\widehat{\Theta}$ et $\widehat{\sigma}^2 = \frac{\|Y-\widehat{Y}\|^2}{n-p-1}$ sont indépendants, et

$$\frac{n-p-1}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-p-1}.$$

Démonstration. C'est encore une conséquence du théorème de Cochran (que l'on peut voir comme une généralisation de la proposition 3), avec la décomposition $\mathbb{R}^n = E \oplus E^\perp$ où $E = \mathrm{Im}(X)$ (de dimension $p+1$) : $\widehat{\Theta}$ ne dépend que de $P_E(Y) = X\widehat{\Theta} = \widehat{Y}$ (puisque X est injective), et $\widehat{\sigma}^2$ ne dépend que de $P_{E^\perp}(Y) = Y - \widehat{Y}$, donc ces variables sont indépendantes. Le second point résulte de $\dim(E^\perp) = n - (p+1)$. □

De cette proposition, on peut déduire intervalle de confiance et tests sur σ^2 (de même que pour la variance en section ? ?), et des régions de confiance pour Θ qui ne dépendent pas de σ^2.
Par exemple, $\widehat{a}_i = \widehat{\Theta}_i$ est un estimateur de a_i et vu la matrice de covariance $\widehat{\Theta}$ on a

$$\widehat{a}_i \sim \mathcal{N}\left(a_i, \sigma^2({}^tXX)^{-1}_{i,i}\right),$$

de sorte que, avec la proposition précédente,

$$\frac{\widehat{a}_i - a_i}{\sqrt{\widehat{\sigma}^2({}^tXX)^{-1}_{i,i}}} \sim t_{n-p-1}$$

d'où l'on déduit intervalle de confiance et tests sur a_i.
Par exemple si la valeur de $\widehat{a}_i$ est proche de 0 (autrement dit, y_k dépend peu de $x_{k,i}$), on peut tester si $\widehat{a}_i = 0$. Plus généralement, tester si certaines composantes de Θ sont nulles s'appelle un test de modèle réduit, et c'est l'objet de la proposition suivante, qui vise à tester si $a_{q+1} = \cdots = a_p = 0$ et comparant les estimations dans les deux cas.

Proposition 7. Soit $0 \leq q \leq p$. On note E_q l'espace vectoriel engendré par les q premières colonnes de la matrice X (de sorte que $E = E_p$) et $\widehat{Y}_q = P_{E_q}(Y)$ (de sorte que $\widehat{Y}_p = \widehat{Y}$). Alors, si $a_{q+1} = \cdots = a_p = 0$,

$$\frac{\|\widehat{Y}_p - \widehat{Y}_q\|^2}{(p-q)\widehat{\sigma}^2} \sim F(p-q, n-p-1).$$

Démonstration. C'est à nouveau une conséquence du théorème de Cochran, appliqué à la décomposition

$$\mathbb{R}^n = E_q \oplus (E_p \cap E_q^\perp) \oplus E_p^\perp.$$

Comme $E_q \subset E_p$, $E_p \cap E_q^\perp$ est l'orthogonal de E_q dans E_p. On a vu que $\widehat{\sigma}^2$ ne dépend que de $P_{E_p^\perp}(Y)$ (et on connaît sa loi). Et la projection de Y sur l'orthogonal de E_q dans E_p (qui est aussi la projection sur $E_q^\perp$ de la projection de Y sur E_p) est $\widehat{Y}_p - \widehat{Y}_q$. Ainsi numérateur et dénominateur de l'énoncé sont indépendants. De plus $\dim E_p \cap E_q^\perp = (p+1) - (q+1) = p - q$. Et $\mathbb{E}[\widehat{Y}_p - \widehat{Y}_q] = X^{(p)}\Theta^{(p)} - X^{(q)}\Theta^{(q)}$ où $X^{(q)}$ est la matrice X réduite à ses $q+1$ premières colonnes, et $\Theta^{(q)}$ est Θ réduit à ses $q+1$ premières lignes (et donc $X^{(p)} = X$ et $\Theta^{(p)} = \Theta$). Par l'hypothèse, on a donc $\mathbb{E}[\widehat{Y}_p - \widehat{Y}_q] = 0$. La conclusion suit, par le théorème de Cochran. □

On peut donc donner un test de $\mathcal{H}_0$: « $a_{q+1} = \cdots = a_n = 0$ » contre $\mathcal{H}_1$ (le contraire), de niveau α, sous la forme : « on rejette l'hypothèse si $\frac{\|\widehat{Y}_p - \widehat{Y}_q\|^2}{(p-q)\widehat{\sigma}^2} > A$ » où A est choisi tel que $\mathbb{P}(F > A) = \alpha$ si $F \sim F(p-q, n-p-1)$. Ce test est consistant quand $p - q \to \infty$.

Lois classiques en statistiques

La **loi du** χ^2 (**khi-deux** ou **chi-deux**) à n degrés de liberté, notée χ^2_n, est la loi de

$$Z_1^2 + \cdots + Z_n^2$$

où $Z_1, \ldots, Z_n$ sont indépendantes et de loi $\mathcal{N}(0,1)$. (NB : son espérance est n)
La **loi de Student** à n degrés de liberté, notée t_n, est la loi de

$$\frac{Z}{\sqrt{U/n}}$$

où $Z \sim \mathcal{N}(0,1)$, $U \sim \chi^2_n$ et ces variables sont indépendante.
La **loi de Fisher-Snedecor** à n et p degrés de liberté, notée $F(n,p)$, est la loi de

$$\frac{U/n}{V/p}$$

où $U \sim \chi^2_n$, $V \sim \chi^2_p$ et ces variables sont indépendantes.