UNIVERSITÉ DE NANTES

Préparation à l'Agrégation

Statistique: compléments de cours

P. Rochet

Contents

1 Résultats préliminaires **3**
1.1 Rappels de probabilités . . . 3
1.2 Vecteurs Gaussiens . . . 4

2 Statistique inférentielle paramétrique **6**
2.1 Modèle d'échantillonage . . . 7
2.2 Méthode des moments . . . 7
2.3 Maximum de vraisemblance . . . 9

3 Estimation non-paramétrique **11**
3.1 Fonction de répartition . . . 11
3.1.1 Fonction de répartition empirique . . . 11
3.1.2 Théorème de Glivenko-Cantelli . . . 13
3.2 Densité . . . 14
3.2.1 Histogramme . . . 14
3.2.2 Histogramme mobile . . . 15
3.2.3 Estimateur à noyaux . . . 17
3.3 Quantiles . . . 19
3.3.1 Fonction de quantile . . . 19
3.3.2 Quantiles empiriques . . . 20

4 Analyse bivariée **22**
4.1 Corrélation . . . 22
4.2 Modélisation non-paramétrique . . . 24

5 Intervalles de confiance **26**
5.1 IC pour la moyenne dans un modèle Gaussien . . . 26
5.2 IC pour la variance dans un modèle Gaussien . . . 27
5.3 IC asymptotiques . . . 27

6 Tests **28**
6.1 Définitions générales 29
6.2 Tests paramétriques 29
6.3 Tests non paramétriques 30
6.3.1 Test sur la moyenne 30
6.3.2 Test du signe 31
6.4 Tests d'adéquation 32
6.4.1 Test du χ^2 32
6.4.2 Test de Kolmogorov-Smirnov 34

7 Modèle linéaire **34**
7.1 Méthode des moindres carrés ordinaires 35
7.2 Comportement asymptotique des estimateurs 37
7.3 Analyse de la variance 38
7.4 Tests 39
7.5 Prediction 40

1 Résultats préliminaires

1.1 Rappels de probabilités

Soit $(X_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires réelles et $(\alpha_n)_{n\in\mathbb{N}}$ une suite réelle strictement positive, on note

- $X_n = o_P(\alpha_n)$ (petit o en probabilité) si X_n/α_n converge en probabilité vers 0.
- $X_n = O_P(\alpha_n)$ (grand o en probabilité) si

$$\forall \delta > 0,\ \exists K > 0,\ \forall n \in \mathbb{N},\ \mathbb{P}\big(|X_n| \leq \alpha_n K\big) \geq 1 - \delta.$$

De manière équivalente, $X_n = O_P(\alpha_n)$ si pour toute suite réelle $(b_n)_{n\in\mathbb{N}}$ qui tend vers 0, $b_n X_n = o_P(\alpha_n)$. Si $X_n = O_P(1)$, on dira que la suite $(X_n)_{n\in\mathbb{N}}$ est bornée en probabilité. On a comme pour les notations $o(.)$ et $O(.)$ déterministes,

$$X_n = o_P(\alpha_n) \iff \frac{X_n}{\alpha_n} = o_P(1) \quad \text{et} \quad X_n = O_P(\alpha_n) \iff \frac{X_n}{\alpha_n} = O_P(1).$$

Remarque 1.1. *Etre borné en probabilité n'implique pas d'être uniformément borné presque sûrement (en revanche, la réciproque est vraie). Comme contre-exemple, on peut prendre une suite iid $(X_n)_{n\in\mathbb{N}}$ de va de loi normale standard qui vérifie $X_n = O_P(1)$ et $Y_n := \max\{X_i : i \leq n\} \xrightarrow{ps} +\infty$.*

On admet les deux théorèmes suivant.

Théorème 1.2 (Slutsky)**.** *Soit $(Y_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires qui converge en loi vers Y et $(Z_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires qui converge en probabilité vers une constante c, alors (Z_n, Y_n) converge en loi vers (c, Y).*

Théorème 1.3 (Lévy)**.** *Soit $(Z_n)_{n\in\mathbb{N}}$ une suite de variables aléatoires dans $\mathbb{R}^k$ de fonctions caratéristiques $\varphi_{Z_n}(t) = \mathbb{E}(e^{i\langle t, Z_n\rangle})$, $t \in \mathbb{R}^k$. Z_n converge en loi vers une variable aléatoire Z si et seulement si φ_{Z_n} converge simplement vers $\varphi_Z : t \mapsto \mathbb{E}(e^{i\langle t, Z\rangle}), t \in \mathbb{R}^k$.*

Les deux théorèmes précédents sont très utiles pour montrer la convergence en loi de variables aléatoires. Ils permettent notamment de montrer facilement la loi faible des grands nombres et le théorème central limite.

Théorème 1.4 (Loi des grands nombres)**.** *Soient $Y_1, ..., Y_n, ...$ des variables aléatoires réelles iid telles que $\mathbb{E}|Y_i| < \infty$, la moyenne empirique $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ converge presque sûrement vers $\mathsf{m} = \mathbb{E}(Y_1)$ quand n tend vers l'infini.*

La convergence presque sûre de la moyenne empirique, qui constitue la version forte de la loi des grands nombres, est un résultat difficile à montrer sans hypothèses supplémentaires. On montrera ici seulement la loi faible des grands nombres, qui établit la convergence en probabilité de $\overline{Y}_n$ vers m. On rappelle que si Y admet des moments finis jusqu'à l'ordre $k \in \mathbb{N}$, m_1,...,m_k, alors

$$\mathbb{E}\big(e^{itY}\big) = \sum_{j=0}^{k} \frac{(it)^j}{j!}\mathsf{m}_j + o(t^k),$$

qui est une conséquence du théorème de convergence dominée.

Preuve de la loi faible des grands nombres. Soit $\varphi_Y : t \mapsto \mathbb{E}(e^{itY_1}), t \in \mathbb{R}$ la fonction caractéristique de Y_1. Comme $\mathbb{E}(Y_1) = m$, on a

$$\varphi_Y(t) = 1 + i\mathsf{m}t + o(t).$$

Par indépendance des Y_i, on a donc

$$\varphi_{\overline{Y}_n}(t) = \mathbb{E}\big(e^{it\overline{Y}_n}\big) = \mathbb{E}\big(e^{i\frac{t}{n}\sum_{i=1}^n Y_i}\big) = \prod_{i=1}^n \mathbb{E}(e^{i\frac{t}{n}Y_i}) = \left[\varphi_Y\Big(\frac{t}{n}\Big)\right]^n = \left[1 + i\mathsf{m}\frac{t}{n} + o\Big(\frac{t}{n}\Big)\right]^n.$$

Quand $n \to \infty$, on a donc $\varphi_{\overline{Y}_n}(t)$ qui converge vers $e^{i\mathsf{m}t}$ pour tout $t \in \mathbb{R}$, qui est la fonction caractéristique de la variable aléatoire constante m. D'après le théorème de Lévy, $\overline{Y}$ converge donc en loi vers m. La convergence en loi vers une constante implique la convergence en probabilité, d'où le résultat. □

1.2 Vecteurs Gaussiens

Le théorème central limite permet d'étudier plus précisément le comportement de la moyenne empirique $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ d'un échantillon iid autour de l'espérance lorsque Y_1 admet un moment d'ordre 2. Sous sa forme vectorielle, la limite est une généralisation vectorielle de la loi normale, appelée vecteur Gaussien. Pour simplifier, on considèrera une constante comme une variable aléatoire Gaussienne de variance nulle.

Définition 1.5. *Soit $\mathsf{m} \in \mathbb{R}^k$ et $\Sigma \in \mathbb{R}^{k\times k}$ une matrice symmétrique semi définie-positive. Un vecteur aléatoire X de $\mathbb{R}^k$ est un vecteur Gaussien de moyenne m et de variance Σ si pour tout vecteur $\mathsf{a} \in \mathbb{R}^k$, $\mathsf{a}^\top X \sim \mathcal{N}(\mathsf{a}^\top\mathsf{m}, \mathsf{a}^\top\Sigma\mathsf{a})$. On le note $X \sim \mathcal{N}(\mathsf{m}, \Sigma)$.*

On vérifie facilement que la matrice Σ est bien la matrice de covariance de X. En effet,

- Pour $\mathsf{a} = \mathsf{e}_i$ le i-ème vecteur de la base canonique, $\mathsf{a}^\top X = X_i \sim \mathcal{N}(\mathsf{m}_i, \Sigma_{ii}) \Longrightarrow \mathrm{var}(X_i) = \Sigma_{ii}$.
- Pour $\mathsf{a} = \mathsf{e}_i + \mathsf{e}_j$, on obtient $X_i + X_j \sim \mathcal{N}(\mathsf{m}_i + \mathsf{m}_j, \Sigma_{ii} + \Sigma_{jj} + 2\Sigma_{ij})$ d'où on déduit $\mathrm{cov}(X_i, X_j) = \Sigma_{ij}$.

Une propriété fondamentale de la famille des vecteurs Gaussiens est qu'elle est stable par transformation affine.

Proposition 1.6. *Soit $X \sim \mathcal{N}(\mathsf{m}, \Sigma)$ un vecteur Gaussien de $\mathbb{R}^k$. Pour $\mathsf{A} \in \mathbb{R}^{p\times k}$ et $\mathsf{b} \in \mathbb{R}^p$, $Y = \mathsf{A}X + \mathsf{b}$ est un vecteur Gaussien de $\mathbb{R}^p$ de loi $Y \sim \mathcal{N}(\mathsf{A}m + \mathsf{b}, \mathsf{A}\Sigma\mathsf{A}^\top)$.*

Preuve. Soit $\mathsf{a} \in \mathbb{R}^p$, on a $\mathsf{a}^\top Y = \mathsf{a}^\top(\mathsf{A}X + \mathsf{b}) = (\mathsf{A}^\top\mathsf{a})^\top X + \mathsf{a}^\top\mathsf{b}$. Comme $X \sim \mathcal{N}(\mathsf{m}, \Sigma)$, on sait que $(\mathsf{A}^\top\mathsf{a})^\top X \sim \mathcal{N}\big((\mathsf{A}^\top\mathsf{a})^\top\mathsf{m}, (\mathsf{A}^\top\mathsf{a})^\top\Sigma(\mathsf{A}^\top\mathsf{a})\big)$. Comme $\mathsf{a}^\top\mathsf{b}$ est une constante, on déduit

$$(\mathsf{A}^\top\mathsf{a})^\top X + \mathsf{a}^\top\mathsf{b} = \mathsf{A}Y + \mathsf{b} \sim \mathcal{N}\big(\mathsf{a}^\top(\mathsf{A}\mathsf{m} + \mathsf{b}), \mathsf{a}^\top(\mathsf{A}\Sigma\mathsf{A}^\top)\mathsf{a}\big).$$

Ceci est valable pour tout $\mathsf{a} \in \mathbb{R}^p$, d'où le résultat. □

Il existe d'autres définitions équivalentes d'un vecteur Gaussien.

Proposition 1.7. *Un vecteur aléatoire X est un vecteur Gaussien de moyenne $\mathsf{m} \in \mathbb{R}^k$ et de variance $\Sigma \in \mathbb{R}^{k\times k}$ si, et seulement si:*

1. *X a pour fonction caractéristique*

$$\varphi_X(t) = \mathbb{E}\big(e^{it^\top X}\big) = \exp\Big(it^\top\mathsf{m} - \frac{t^\top\Sigma t}{2}\Big)\ ,\ t \in \mathbb{R}^k.$$

2. *Si Σ est inversible, X a pour densité*

$$f_X(x) = \frac{1}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}}\exp\Big(-\frac{1}{2}(x-\mathsf{m})^\top\Sigma^{-1}(x-\mathsf{m})\Big)\ ,\ x \in \mathbb{R}^k.$$

Preuve. Soit X qui vérifie la définition 1.5, on remarque que

$$\varphi_X(t) = \varphi_{t^\top X}(1) = \exp\Big(i\ t^\top \mathsf{m} - \frac{t^\top \Sigma t}{2}\Big),$$

où le terme de droite est une simple évaluation de la fonction caractéristique d'une variable aléatoire Gaussienne réelle $\mathcal{N}(t^\top \mathsf{m}, t^\top \Sigma t)$. De plus, on vérifie facilement qu'un vecteur aléatoire de densité f_X a pour fonction caractéristique φ_X. Le résultat découle du fait que la fonction caractéristique caractérise la loi. □

Remarque 1.8. *Lorsque la matrice Σ n'est pas inversible, X "vit" dans le sous-espace affine $\mathcal{A} = \{\mathsf{m}\} \oplus \mathrm{Im}(\Sigma) \subset \mathbb{R}^k$, dans le sens où $\mathbb{P}(X \in \mathcal{A}) = 1$. Dans ce cas, X n'admet pas de densité sur $\mathbb{R}^k$ par rapport à la mesure de Lebesgue.*

Proposition 1.9. *Soit $X \sim \mathcal{N}(\mathsf{m}, \Sigma)$ et $Y_1 = \mathsf{A}_1 X + \mathsf{b}_1$, $Y_2 = \mathsf{A}_2 X + \mathsf{b}_2$ deux transformations affines de X à valeurs dans $\mathbb{R}^{p_1}$ et $\mathbb{R}^{p_2}$ respectivement. Les énoncés suivants sont équivalents:*

i) Y_1 *et* Y_2 *sont indépendants*

ii) Y_1 *et* Y_2 *sont non-corrélés:* $\mathbb{E}(Y_1 Y_2^\top) - \mathbb{E}(Y_1)\mathbb{E}(Y_2^\top) = 0 \in \mathbb{R}^{p_1 \times p_2}$

iii) $\mathsf{A}_1 \Sigma \mathsf{A}_2^\top = 0 \in \mathbb{R}^{p_1 \times p_2}$

En particulier, les coordonnées d'un vecteur Gaussien sont indépendantes si et seulement si elles sont non-corrélées.

Preuve. Clairement *i)* $\implies$ *ii)*. De plus, $\mathbb{E}(Y_1 Y_2^\top) - \mathbb{E}(Y_1)\mathbb{E}(Y_2^\top) = \mathsf{A}_1 \Sigma \mathsf{A}_2^\top$ d'où *ii)* $\implies$ *iii)*. Enfin, $Y := (Y_1^\top, Y_2^\top)^\top$ est un vecteur Gaussien par la proposition 1.6. Pour $t = (t_1, t_2)^\top \in \mathbb{R}^{p_1+p_2}$, sa fonction caractéristique est donnée par

$$\varphi_Y(t) = \exp\Big(i\ (t_1^\top \mathsf{b}_1 + t_2^\top \mathsf{b}_2) - \frac{t_1^\top \mathsf{A}_1 \Sigma \mathsf{A}_1^\top t_1 + t_2^\top \mathsf{A}_1 \Sigma \mathsf{A}_1^\top t_1 + 2 t_1^\top \mathsf{A}_1 \Sigma \mathsf{A}_2^\top t_2}{2}\Big).$$

On a donc $\varphi_Y(t) = \varphi_{Y_1}(t_1)\varphi_{Y_2}(t_2)$ si et seulement si $\mathsf{A}_1 \Sigma \mathsf{A}_2^\top = 0$. □

On est maintenant en mesure de présenter la version vectorielle du théorème central limite.

Théorème 1.10 (Théorème central limite)**.** *Soient $Y_1, ..., Y_n, ...$ des variables aléatoires de $\mathbb{R}^k$ iid de carré intégrable telles que $\mathbb{E}(Y_i) = \mathsf{m} \in \mathbb{R}^k$ et $\mathrm{var}(Y_i) = \Sigma \in \mathbb{R}^{k \times k}$, l'estimateur $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ vérifie*

$$\sqrt{n}(\overline{Y}_n - \mathsf{m}) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Sigma).$$

La preuve du théorème central limite utilise également le théorème de Lévy, en développant à l'ordre 2 la fonction caractéristique au voisinage de zéro.

Preuve du théorème central limite. On pose $Z_i = Y_i - \mathsf{m}$. On a $\mathbb{E}(Z_i) = 0$, $\mathbb{E}(Z_i Z_i^\top) = \Sigma$ et

$$\varphi_Z(t) = \mathbb{E}(e^{i\langle t, Z_1\rangle}) = 1 - \frac{t^\top \Sigma t}{2} + o(\|t\|^2).$$

En écrivant $\sqrt{n}(\overline{Y}_n - \mathsf{m}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i$, on déduit

$$\varphi_{\sqrt{n}(\overline{Y}_n - \mathsf{m})}(t) = \mathbb{E}\big(e^{i\langle t, \sqrt{n}(\overline{Y}_n - \mathsf{m})\rangle}\big) = \prod_{i=1}^n \mathbb{E}\big(e^{i\frac{1}{\sqrt{n}}\langle t, Z_i\rangle}\big) = \Big[\varphi_Z\Big(\frac{t}{\sqrt{n}}\Big)\Big]^n = \Big[1 - \frac{t^\top \Sigma t}{2n} + o\Big(\frac{\|t\|^2}{n}\Big)\Big]^n.$$

Pour tout $t \in \mathbb{R}^k$, la fonction caractéristique de $\sqrt{n}(\overline{Y}_n - \mathsf{m})$ au point t converge donc vers $e^{-\frac{1}{2}t^\top \Sigma t}$, la fonction caractéristique d'une loi normale $\mathcal{N}(0, \Sigma)$. On conclut par le théorème de Lévy. □

Définition 1.11. *La loi du Khi-2 à k degrés de liberté, notée $\chi^2(k)$, est la loi de la somme des carrés de k variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$.*

Le théorème suivant concerne la loi de projections orthogonales de vecteurs Gaussiens.

Théorème 1.12 (Théorème de Cochran)**.** *Soit X un vecteur Gaussien standard de $\mathbb{R}^k$, $X \sim \mathcal{N}(0, \mathrm{I})$. Soit E un sous-espace vectoriel de $\mathbb{R}^k$ de dimension d et Π_E sa matrice de projection orthogonale. Alors, $\Pi_E X$ et $(\mathrm{I} - \Pi_E)X$ sont des vecteurs Gaussiens indépendants de lois $\mathcal{N}(0, \Pi_E)$ et $\mathcal{N}(0, \mathrm{I} - \Pi_E)$ respectivement. De plus, $\|\Pi_E X\|^2$ et $\|(\mathrm{I} - \Pi_E)X\|^2$ sont des variables aléatoires indépendantes de lois $\chi^2(d)$ et $\chi^2(n-d)$.*

Preuve. Les lois de $\Pi_E X$ et $(\mathrm{I} - \Pi_E)X$ de déduisent immédiatement de la Proposition 1.6. De plus, la fonction caractéristique de $X = \Pi_E X + (\mathrm{I} - \Pi_E)X$ est égale au produit des fonctions caractéristiques de $\Pi_E X$ et $(\mathrm{I} - \Pi_E)X$ par le théorème de Pythagore, les vecteurs $\Pi_E X$ et $(\mathrm{I} - \Pi_E)X$ sont donc indépendants. Soit $\mathsf{u}_1, ..., \mathsf{u}_d$ une base orthonormée de E et $\mathsf{U} = (\mathsf{u}_1, ..., \mathsf{u}_d) \in \mathbb{R}^{k \times d}$. On remarque que $\mathsf{U}^\top \mathsf{U} = \mathrm{I}$ (la matrice identité de $\mathbb{R}^d$) et $\mathsf{U}\mathsf{U}^\top = \Pi_E$. Par la proposition 1.6, $\mathsf{U}^\top X$ est un vecteur Gaussien standard de $\mathbb{R}^d$. D'où

$$\|\Pi_E X\|^2 = \|\mathsf{U}\mathsf{U}^\top X\|^2 = \|\mathsf{U}^\top X\|^2 = \sum_{i=1}^{d} (\mathsf{u}_i^\top X)^2 \sim \chi^2(d).$$

La loi de $\|(\mathrm{I} - \Pi_E)X\|^2$ s'obtient de la même manière. □

Le théorème de Cochran est notamment utilisé pour construire les tests de significativité des coefficients dans un modèle de régression linéaire Gaussien.

2 Statistique inférentielle paramétrique

Contrairement aux statistiques descriptives qui consistent essentiellement à résumer l'information contenue dans des données à l'aide d'indicateurs (moyenne, médiane, variance etc...) ou de représentations graphiques (diagrammes, boxplots, histogrammes etc...), la statistique inférentielle cherche à fournir un modèle probabiliste qui explique les données observées. Les observations sont donc traitées comme des réalisations de variables aléatoires de loi inconnue, dont on cherche à retrouver certaines caractéristiques (ex: espérance, probabilité d'un événement, densité etc...).

La statistique inférentielle permet de répondre à différentes questions:

1. Le choix d'un modèle permettant de décrire le phénomène aléatoire
2. L'estimation de la loi des observations
3. La vérification d'hypothèses par des tests

Plus le modèle est précis, plus il sera facile d'évaluer la loi des observations. Par exemple, si on observe un échantillon $x_1, ..., x_n$ issu d'une loi normale $\mathcal{N}(m, \sigma^2)$, il suffit d'estimer la moyenne et la variance des observations pour connaître la loi. En revanche, on prend le risque de considérer un modèle qui n'est pas exact. Si les observations ne sont pas Gaussiennes, on ne retrouvera jamais leur loi en se limitant à un modèle Gaussien.

Définition 2.1. *Un modèle paramétrique est un modèle pour lequel la loi des observations dépend d'une ou plusieurs constantes inconnues appelées paramètres.*

Par exemple, le modèle Gaussien $\{\mathcal{N}(m, \sigma^2), m \in \mathbb{R}, \sigma^2 > 0\}$, le modèle de Poisson $\{\mathcal{P}(\lambda), \lambda > 0\}$ ou le modèle exponentiel $\{\mathcal{E}(\lambda), \lambda > 0\}$ sont des modèles paramétriques.

Définition 2.2. *Un modèle non-paramétrique est un modèle qui ne peut pas être décrit par un nombre fini de paramètres.*

Utiliser un modèle non-paramétrique signifie généralement qu'on dispose de peu d'information sur la loi μ des observations. Par exemple, μ est une loi continue (densité), μ admet un moment d'ordre deux, μ est une loi symétrique par rapport à zéro ou encore μ est une loi de probabilité (aucune information) sont des exemples de modélisation non-paramétriques.

Lorsqu'on choisit un modèle paramétrique, on prend le risque que la vraie loi des données n'appartienne pas au modèle. Un modèle non-paramétrique permet donc une évaluation parfois moins précise mais plus robuste de cette loi.

2.1 Modèle d'échantillonage

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé et $\mathcal{X}$ un sous-ensemble de $\mathbb{R}$. On observe une réalisation $(x_1, ..., x_n) \in \mathcal{X}^n$ d'un vecteur aléatoire $(X_1, ..., X_n)$. On suppose les composantes $X_1, ..., X_n$ indépendantes et de même loi μ sur $\mathcal{X}$. On dira que le vecteur $(X_1, ..., X_n)$ est un échantillon de variables indépendantes et identiquement distribuées, abrégé en iid.

Il faut distinguer l'échantillon $(X_1, ..., X_n)$ qui est aléatoire, et les données $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$ qui sont la réalisation d'un événement ω. Lors d'une étude statistique, on observe seulement l'image de ω par les fonctions $X_1, ..., X_n$.

Définition 2.3. *Une statistique $T = T(X_1, ..., X_n)$ est une fonction connue de l'échantillon.*

Une statistique T est une variable aléatoire dont on observe une réalisation $T(x_1, ..., x_n)$. On parle d'estimation lorsque l'on cherche à approcher une caractéristique de la loi μ (par exemple sa moyenne) par une fonction des observations. La statistique utilisée est appelée *estimateur*.

Définition 2.4. *Le biais d'un estimateur T de θ est la quantité $b(T) = \mathbb{E}(T) - \theta$.*

Le biais n'est défini que si l'espérance de T est définie, c'est-à-dire si $\mathbb{E}|T| < \infty$. Si le biais est nul, on dira que T est un estimateur sans biais, ce qui est évidemment une propriété appréciable. Cela signifie que l'estimateur est égal "en moyenne" au paramètre à estimer.

Définition 2.5. *L'erreur quadratique moyenne (EQM) d'un estimateur T de θ est la quantité*

$$\mathrm{EQM}(T) = \mathbb{E}(T - \theta)^2 \in [0, +\infty].$$

On dit qu'un estimateur T de θ converge dans $\mathbb{L}^2$ si son EQM tend vers zéro. La formule de la décomposition biais-variance:

$$\mathrm{EQM}(T) = b(T)^2 + \mathrm{var}(T)$$

permet de montrer immédiatement qu'un estimateur converge dans $\mathbb{L}^2$ si et seulement si son biais et sa variance tendent vers zéro.

2.2 Méthode des moments

Une approche classique en statistique consiste à utiliser comme approximation de la loi μ inconnue, la loi empirique de l'échantillon

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_n},$$

où δ représente la mesure de Dirac. La mesure de probabilité aléatoire μ_n, construite à partir de l'échantillon, converge en un sens faible vers la vraie loi des observations, puisque l'on sait que, en conséquence de la loi forte de grands nombres,

$$\int g(x)d\mu_n(x) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p.s.} \mathbb{E}(g(X)),$$

pour toute fonction g telle que $\int |g(x)|d\mu(x) < \infty$. Ainsi, la loi empirique s'avère être un outil intéressant pour estimer des paramètres linéaires, de la forme

$$\theta = \int g(x)d\mu(x).$$

La méthode des moments consiste à remplacer la loi μ des observations (inconnue) par la loi empirique μ_n pour estimer les paramètres linéaires: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$.

Proposition 2.6. *Soit g une fonction mesurable telle que $\int |g|d\mu < \infty$ et $\theta = \int g d\mu$. L'estimateur de θ par la méthode des moments:*

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

est un estimateur sans biais de θ qui converge presque sûrement quand $n \to \infty$.

Preuve. On vérifie facilement que $\hat{\theta}$ est sans biais. La convergence presque sûre est une conséquence de la loi forte des grands nombres. □

Remarque 2.7. *Le plus souvent, la méthode des moments fait référence à une méthode d'estimation de moments d'une variable aléatoire X, définis par*

$$\mathbb{E}(X^p),\ p = 1, 2, ...$$

Nous appelons ici méthode des moments une généralisation qui sétend à tout paramètre linéaire.

Exemples

1. Moyenne empirique: l'estimateur

$$\hat{\mathsf{m}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

est un estimateur sans biais de la moyenne $\mathsf{m} = \mathbb{E}(X)$. Cela correspond à la fonction $g(x) = x$.

2. Moments d'ordre p: si $\mathbb{E}|X|^p < \infty$, alors

$$\hat{\mathsf{m}}_p = \frac{1}{n}\sum_{i=1}^{n} X_i^p$$

est un estimateur sans biais du moment d'ordre p, $\mathsf{m}_p = \mathbb{E}(X^p)$. Cela correspond à $g(x) = x^p$.

3. Fonction de répartition empirique: un paramètre de la forme $\theta = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} d\mu(x)$ peut être estimé par son homologue empirique

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}.$$

Cela correspond à $g(x) = \mathbb{1}\{x \leq t\}$.

Proposition 2.8. *Soient $g_1, ..., g_k$ une collection de fonctions à valeurs dans $\mathbb{R}$ telles que $\int |g_j(x)| d\mu < \infty$, on note $\mathsf{m}_j = \int g_j d\mu$ et $\hat{\mathsf{m}}_j$ l'estimateur de m_j par la méthode des moments. Soit $G : \mathbb{R}^k \to \mathbb{R}$ une fonction continue, l'estimateur*

$$\hat{\theta} = G\left(\hat{\mathsf{m}}_1, ..., \hat{\mathsf{m}}_k\right),$$

converge presque sûrement vers $\theta = G(\mathsf{m}_1, ..., \mathsf{m}_k)$.

Preuve. Par la LFGN, $\hat{\mathsf{m}}_j$ converge presque sûrement vers m_j pour tout $j = 1, ..., k$. Comme une intersection finie d'événements de probabilité 1 est de probabilité 1, le vecteur $(\mathsf{m}_1, ..., \mathsf{m}_k)$ converge presque sûrement. Par continuité de G (la continuité en $(\mathsf{m}_1, ..., \mathsf{m}_k)$ suffit), $\hat{\theta} \xrightarrow{ps} \theta$. □

Contrairement au cas où le paramètre à estimer est linéaire, l'estimateur $\hat{\theta}$ n'est généralement pas sans biais. Ce résultat nous montre que la méthode des moments fournit un moyen d'estimer tout paramètre qui peut s'écrire comme une fonction de plusieurs moments de μ. Un exemple classique est l'estimateur de la variance d'un échantillon. Supposons que $\mathbb{E}(X^2) < \infty$, alors l'estimateur

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2,$$

converge presque sûrement vers $\sigma^2 = \mathrm{var}(X)$ en conséquence de la proposition 2.8. Cela correspond au cas $g_1(x) = x^2$, $g_2(x) = x$ et $G(a, b) = a - b^2$. On peut montrer par ailleurs que $\hat{\sigma}^2$ n'est pas sans biais mais est asymptotiquement sans biais.

Applications

1. Modèle de Poisson: on observe un échantillon iid de loi de Poisson de paramètre λ inconnu. On sait que l'espérance et la variance d'une loi de Poisson sont égales au paramètre. Donc, on peut construire deux estimateurs différents, à savoir,

$$\hat{\lambda}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \hat{\lambda}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

   On remarque que $\hat{\lambda}^{(1)}$ est sans biais, contrairement à $\hat{\lambda}^{(2)}$. De plus, on peut montrer que $\mathrm{EQM}(\hat{\lambda}^{(1)}) < \mathrm{EQM}(\hat{\lambda}^{(2)})$ ce qui conduit à la conclusion que l'estimateur $\hat{\lambda}^{(1)}$ est meilleur au sens $\mathbb{L}^2$.

2. Loi exponentielle: on observe un échantillon iid de loi exponentielle de paramètre λ inconnu. L'espérance d'une loi exponentielle est égale à $1/\lambda$, on peut donc estimer λ par

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i}$$

   qui n'est pas sans biais.

2.3 Maximum de vraisemblance

Dans un problème d'estimation paramétrique, une notion différente de la méthode des moments, mais très utilisée est le *maximum de vraisemblance*. L'idée est de choisir comme estimateur la valeur du paramètre qui rend la plus probable l'observation de l'échantillon $x_1, ..., x_n$.

Un modèle paramétrique $\{\mu_\theta : \theta \in \Theta\}$ est dit *identifié* si l'application $\theta \mapsto \mu_\theta$ est injective sur Θ. Autrement dit, un modèle est identifié si chaque valeur du paramètre est associé à une unique loi du modèle.

Définition 2.9. *Soit $\mathcal{M} = \{\mu_\theta, \theta \in \Theta\}$ un modèle paramétrique identifié et $x_1, ..., x_n$ les observations.*

1. *Si pour tout $\theta \in \Theta$, μ_θ est une loi discrète, la vraisemblance du modèle est la fonction*

$$\theta \mapsto L(\theta, x_1, ..., x_n) = \prod_{i=1}^{n} \mu_\theta(\{x_i\}),\ \theta \in \Theta.$$

2. *Si pour tout $\theta \in \Theta$, μ_θ est une loi continue de densité f_θ, la vraisemblance du modèle est la fonction*

$$\theta \mapsto L(\theta, x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i),\ \theta \in \Theta.$$

Un estimateur du maximum de vraisemblance est une variable aléatoire qui maximise $\theta \mapsto L(\theta, X_1, ..., X_n)$.

Exemples

1. Modèle binomial: on observe le résultat de n lancés de dés $(x_1, ..., x_n) \in \{0, 1\}^n$. Les x_i sont des réalisations indépendantes d'une variable de Bernoulli de paramètre p_0 inconnu. La probabilité d'observer l'échantillon est de

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} p_0^{x_i}(1 - p_0)^{1-x_i} = p_0^{\sum x_i}(1 - p_0)^{n-\sum x_i}.$$

La vraie valeur p_0 étant inconnue, on peut écrire cette probabilité comme une fonction de p,

$$L(p, x_1, ..., x_n) = p^{\sum x_i}(1 - p)^{n-\sum x_i}$$

qui est la probabilité d'observer l'échantillon $(x_1, ..., x_n)$ si les observations suivent la loi de Bernoulli de paramètre p (qui n'est autre que la vraisemblance du modèle). L'estimateur du maximum de vraisemblance est la valeur $\hat{p}$ qui maximise la probabilité $L(p, x_1, ..., x_n)$,

$$\hat{p} = \arg\max_{p \in [0,1]} L(p, x_1, ..., x_n).$$

On note $\ell(p, x_1, ..., x_n) = \log L(p, x_1, ..., x_n)$ la log-vraisemblance, maximiser $L(., x_1, ..., x_n)$ revient à maximiser $\ell(., x_1, ..., x_n)$. La fonction $p \mapsto \ell(p, x_1, ..., x_n)$ est strictement concave sur $[0, 1]$, le maximiseur est donc l'unique solution de la condition du premier ordre

$$\frac{\partial}{\partial\theta}\ \ell(\hat{p}, x_1, ..., x_n) = \frac{\sum x_i}{\hat{p}} - \frac{n - \sum x_i}{1 - \hat{p}} = 0 \Longleftrightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Dans ce modèle l'estimateur du maximum de vraisemblance est donc défini comme la variable aléatoire $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2. Modèle de Poisson: la vraisemblance pour un modèle de Poisson $\mathcal{M} = \{\mathcal{P}(\lambda), \lambda > 0\}$ est donnée par

$$L(\lambda, x_1, ..., x_n) = e^{-n\lambda}\prod_{i=1}^{n}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\Pi\ x_i!}.$$

Maximiser $L(\lambda, x_1, ..., x_n)$ revient à maximiser $\lambda \mapsto \log(\lambda)\sum_{i=1}^{n} x_i - n\lambda$. Le maximiseur est

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

3. Loi uniforme: on considère le modèle $\mathcal{M} = \{\mathcal{U}(0,\theta), \theta > 0\}$, qui sont des lois de densités

$$f_\theta(x) = \begin{cases} 1/\theta & \text{si } x \in [0,\theta] \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance du modèle est donnée par

$$L(\theta, x_1, ..., x_n) = \begin{cases} \theta^{-n} & \text{si } \max\{x_1, ..., x_n\} \leq \theta \\ 0 & \text{sinon.} \end{cases}$$

La fonction $\theta \mapsto \theta^{-n}$ est décroissante sur $[0, +\infty[$, on déduit donc que la vraisemblance est maximale pour $\hat{\theta} = \max\{x_1, ..., x_n\}$.

4. Loi géométrique: la vraisemblance pour un modèle géométrique $\mathcal{M} = \{\mathcal{G}(p), p \in [0,1]\}$ est donnée par

$$L(p, x_1, ..., x_n) = p^n (1-p)^{\sum x_i},$$

qui est maximale pour $\hat{p} = n / \sum_{i=1}^n x_i$.

Le maximum de vraisemblance est invariant par reparamétrisation du modèle.

Proposition 2.10. *Soit $\hat{\theta}$ un estimateur du maximum de vraisemblance d'un paramètre θ et ϕ une fonction injective définie sur Θ, $\phi(\hat{\theta})$ est un estimateur du maximum de vraisemblance de $\phi(\theta)$.*

Preuve. Soit $L(\theta, x_1, ..., X_n)$ la vraisemblance du modèle. La vraisemblance du modèle reparamétré par $\phi(\theta)$ est l'application définie sur $\phi(\Theta)$ par $\eta \mapsto L(\phi^{-1}(\eta), x_1, ..., x_n)$, qui atteint son maximum en $\eta = \phi(\hat{\theta})$. □

Une reparamétrisation d'un modèle paramétrique identifié peut le rendre non identifié si la fonction de reparamétrisation ϕ n'est pas injective. C'est évidemment à éviter, mais dans ce cas, on peut définir la vraisemblance par

$$\eta \mapsto \sup_{\theta : \phi(\theta) = \eta} L(\theta, x_1, ..., x_n).$$

3 Estimation non-paramétrique

On observe la réalisation d'un échantillon $X_1, ..., X_n$ de variable aléatoires iid de loi μ. L'estimation *non-paramétrique* désigne l'estimation de caractéristiques qui appartiennent à un espace de dimension infinie. Dans ce chapitre, on s'intéressera en particulier à l'estimation de la fonction de répartition, de la densité et de la fonction de quantile.

3.1 Fonction de répartition

3.1.1 Fonction de répartition empirique

Soit $X_1, ..., X_n$ un échantillon de variables aléatoires iid réelles ($\mathcal{X} \subseteq \mathbb{R}$) de loi μ.

Définition 3.1. *La fonction de répartition de μ est donnée par*

$$F(x) = \mu(]-\infty, x]) = \mathbb{P}(X \leq x), \ x \in \mathbb{R}.$$

La fonction de répartition caractérise la loi, c'est-à-dire que deux lois μ et ν sont égales si, et seulement si, leurs fonctions de répartition sont égales. On rappelle qu'une fonction de répartition est une fonction croissante et telle que $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$. De plus, F est continue à droite et limitée à gauche (càdlàg) avec

$$\forall x \in \mathbb{R}, \ \lim_{y\to x^+} F(y) = F(x) \quad \text{et} \quad \lim_{y\to x^-} F(y) = F(x^-) = \mathbb{P}(X < x).$$

On a donc en particulier $\mathbb{P}(X = x) = F(x) - F(x^-)$. Par ailleurs, lorsque μ admet une densité f continue, la fonction de répartition vaut

$$F(x) = \int_{-\infty}^{x} f(y)dy, \ x \in \mathbb{R}.$$

Dans ce cas F est de classe $\mathcal{C}^1$ de dérivée $F' = f$ par le théorème fondamental de l'analyse.

Pour estimer la fonction de répartition en un point x, on utilise la proportion des observations inférieures à x,

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}, \ x \in \mathbb{R},$$

où $\mathbb{1}$ est la fonction indicatrice. La fonction F_n est appelée *fonction de répartition empirique.*

Proposition 3.2. *Pour tout $\omega \in \Omega$, l'estimation $F_n : \mathbb{R} \to [0, 1]$ est une fonction de répartition.*

Preuve. Soit $\omega \in \Omega$, on pose $(x_1, ..., x_n) = (X_1(\omega), ..., x_n(\omega))$. La fonction de répartition empirique associée est

$$\big(F_n(x)\big)(\omega) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}, \ x \in \mathbb{R}.$$

Montrons qu'elle vérifie les propriétés d'une fonction de répartition. Pour tout $x \geq \max\{x_1, ..., x_n\}$, $F_n(x) = 1$. De même, pour tout $x < \min\{x_1, ..., x_n\}$, $F_n(x) = 0$. On a donc clairement

$$\lim_{x\to-\infty} F_n(x) = 0 \quad \text{et} \quad \lim_{x\to+\infty} F_n(x) = 1.$$

Pour tout i, la fonction $x \mapsto \mathbb{1}\{x_i \leq x\}$ est croissante, continue sur $\mathbb{R} \setminus \{x_i\}$ et càdlàg en x_i. On déduit que F_n est croissante et càdlàg comme somme finie de fonctions croissantes càdlàg. On a précisemment en x_i,

$$\lim_{x\to x_i^+} F_n(x) = \frac{\#\{j : x_j \leq x_i\}}{n} = F_n(x_i)$$

et F_n admet une limite à gauche

$$\lim_{x\to x_i^-} F_n(x) = \frac{\#\{j : x_j < x_i\}}{n} = F_n(x_i) - \frac{\#\{j : x_j = x_i\}}{n}. \qquad \square$$

Remarque 3.3. *L'ensemble $\{j : x_j = x_i\}$ admet au moins un élément: x_i. La fonction de répartition empirique F_n n'est donc jamais continue en x_i puisque la limite à gauche est strictement inférieure à la limite à droite.*

Lorsque tous les x_i sont différents (ce qui arrive presque sûrement lorsqu'ils sont de loi continue), la fonction de répartition empirique F_n obtenue pour une réalisation ω est la fonction de répartition de la loi uniforme discrète sur $\{x_1, ..., x_n\}$, parfois appelée loi empirique. On approche donc en quelque sorte la loi des observations par une loi discrète sur l'ensemble des réalisations. En particulier, $F_n(x)$ ne peut prendre que $n + 1$ valeurs différentes: $0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1$.

Proposition 3.4. *Pour tout $x \in \mathbb{R}$, $F_n(x)$ est un estimateur sans biais de $F(x)$ qui converge presque sûrement vers $F(x)$ quand $n \to \infty$.*

Preuve. Soit $Y_i = \mathbb{1}\{X_i \leq x\}$, les variables aléatoires $Y_1, ..., Y_n$ sont indépendantes de loi de Bernoulli $\mathcal{B}(F(x))$. On a donc en particulier, $\mathbb{E}(Y_i) = F(x)$ et $\mathrm{var}(Y_i) = F(x)(1 - F(x)) < \infty$. On déduit

$$\mathbb{E}(F_n(x)) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i\Big) = \mathbb{E}\left(Y_1\right) = F(x).$$

Par la loi forte des grands nombres, $F_n(x)$ converge presque sûrement vers $F(x)$. □

3.1.2 Théorème de Glivenko-Cantelli

La loi forte des grands nombres permet d'établir facilement la convergence presque sûre de la fonction de répartition empirique vers la vraie fonction de répartition en tout point. Le théorème de Glivenko-Cantelli donne un résultat plus fort, à savoir, la convergence uniforme presque sûre de la fonction de répartition empirique.

Théorème 3.5 (Glivenko-Cantelli)**.** *Soit $X_1, .., X_n$ un échantillon iid de fonction de répartition continue F. La fonction de répartition empirique F_n converge presque sûrement uniformément vers F quand $n \to \infty$:*

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \to \infty]{p.s.} 0.$$

Pour la preuve, on aura besoin du lemme suivant.

Lemme 3.6. *Si X est une variable aléatoire réelle de fonction de répartition F, alors pour tout $x \in \mathbb{R}$,*

$$\mathbb{1}\{X \leq x\} \overset{ps}{=} \mathbb{1}\{F(X) \leq F(x)\}.$$

En particulier, si F est continue, alors $U := F(X)$ est de loi uniforme sur $]0, 1[$.

Preuve du lemme. Comme F est croissante, on sait que

$$\{F(X) \leq F(x)\} = \{X \leq x\} \cup \{F(X) \leq F(x), x < X\}.$$

Soit $u = \sup\{y : F(y) \leq F(x)\}$, on distingue deux cas. Si $F(u) = F(x)$ alors

$$\mathbb{P}(F(X) \leq F(x), x < X) = \mathbb{P}(F(X) \leq F(x), x < X \leq u) \leq F(u) - F(x) = 0.$$

Sinon, soit u_n une suite croissante qui tend vers u et telle que $F(u_n) = F(x)$ pour tout $n \in \mathbb{N}$. On a

$$\begin{aligned}\mathbb{P}(F(X) \leq F(x), x < X) &= \lim_{n \to \infty} \mathbb{P}(F(X) \leq F(x), x < X \leq u_n) \\ &\leq \lim_{n \to \infty} F(u_n) - F(x) = 0,\end{aligned}$$

où la première égalité découle de l'axiome des probabilité $\lim_{n \to \infty} \mathbb{P}(\cup_{k=1}^{n} A_k) = \mathbb{P}(\cup_{n=1}^{\infty} A_n)$. On déduit $\mathbb{1}\{X \leq x\} \overset{ps}{=} \mathbb{1}\{F(X) \leq F(x)\}$. Si F est continue, du fait que $\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to \infty} F(x) = 1$, il existe, pour tout $u \in]0, 1[$, un (ou plusieurs) $x \in \mathbb{R}$ tel que $F(x) = u$. Par conséquent, l'égalité $\mathbb{P}(F(X) \leq u) = u$ est vérifiée pour tout $u \in]0, 1[$, ce qui prouve que $F(X)$ est de loi uniforme sur $]0, 1[$. □

La preuve du théorème de Glivenko-Cantelli utilise également le théorème suivant, dû à Polyà (parfois appelé théorème de Dini) que l'on admettra.

Théorème 3.7. *Une suite F_n de fonctions croissantes définies sur un compact $[a,b] \subset \mathbb{R}$ qui converge simplement vers une fonction F continue sur $[a,b]$, converge uniformément vers F.*

Preuve du théorème de Glivenko-Cantelli. D'après le lemme 3.6, on a

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \le x\} - F(x) \right| \overset{ps}{=} \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{F(X_i) \le F(x)\} - F(x) \right|.$$

Par hypothèse, on sait que $F(x)$ peut prendre toutes les valeurs de l'intervalle $]0,1[$. En posant $u = F(x)$ et $U_i = F(X_i)$ qui suit une loi uniforme sur $[0,1]$, on a donc

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \overset{ps}{=} \sup_{u\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{U_i \le u\} - u \right|,$$

les valeurs en $u = 0$ et $u = 1$ étant trivialement 0. Il suffit de montrer le résultat pour un échantillon $U_1, ..., U_n$ de variables aléatoires uniformes sur $[0,1]$. On suppose donc sans perte de généralité que

$$F_n(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{U_i \le u\} \quad \text{et} \quad F(u) = u,\ u \in [0,1].$$

Soit $q \in \mathbb{Q}\cap]0,1[$, on sait par la proposition 3.4 que $\mathbb{P}(F_n(q) \to q) = 1$. Comme une union dénombrable d'ensembles de probabilité nulle est de probabilité nulle, on déduit que l'événement $\Omega_0 := \{\forall q \in \mathbb{Q}\cap]0,1[, F_n(q) \to F(q)\}$ est de probabilité 1. Soit $u \in]0,1[$ et $(u_k)_{k\in\mathbb{N}}$ et $(v_k)_{k\in\mathbb{N}}$ deux suites rationnelles adjacentes de limite u telles que $u_k \le u \le v_k,\ \forall k \in \mathbb{N}$. La croissance de F_n entraîne

$$\forall k \in \mathbb{N},\ F_n(u_k) \le F_n(u) \le F_n(v_k).$$

En faisant tendre n vers l'infini, on déduit que sur Ω_0,

$$F(u_k) \le \liminf_{n\to\infty} F_n(u) \le \limsup_{n\to\infty} F_n(u) \le F(v_k).$$

Comme le résultat est vrai pour tout k, il reste vrai en faisant tendre k vers l'infini. On sait de plus que $F_n(0) = 0$ et $F_n(1) = 1$ presque sûrement. On obtient $\Omega_0 = \{\forall u \in [0,1], F_n(u) \to u\}$. Finalement, d'après le théorème 3.7, F_n converge uniformément vers F sur Ω_0. □

Remarque 3.8. *Le théorème de Glivenko-Cantelli reste vrai lorsque F n'est pas continue mais la preuve précédente n'est plus valable dans ce cas.*

On peut montrer que la loi de $K_n := \sqrt{n}\ \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ ne dépend pas de F lorsque celle-ci est continue. Cette loi dépend de n mais converge vers la loi de Kolmogorov quand $n \to \infty$. Ce résultat qui est une conséquence du théorème de Donsker, est notamment utile pour la construction du test de Kolmogorov-Smirnov.

3.2 Densité

3.2.1 Histogramme

La méthode la plus simple pour estimer une densité à partir de données est par un *histogramme*. C'est souvent la méthode privilégiée pour représenter graphiquement les données. On se fixe une famille ordonnée $a_0 < a_1 < ... < a_k$ de réels tels que $a_0 \le \min\{x_1, ..., x_n\}$ et $a_k > \max\{x_1, ..., x_n\}$. L'histogramme associé à la suite $(a_j)_{j=0,...,k}$ est la fonction F_n constante sur les intervalles (ou classes) $[a_{j-1}, a_j[$, $j =$

$1, ..., k$, qui associe à x la proportion des observations appartenant au même intervalle que x, renormalisée par la taille de l'intervalle. Formellement, F_n est définie par

$$F_n(x) = \frac{1}{a_j - a_{j-1}} \sum_{i=1}^{n} \frac{\mathbb{1}\{a_{j-1} \leq x_i < a_j\}}{n} = \frac{\#\{i : x_i \in [a_{j-1}, a_j[\}}{n(a_j - a_{j-1})}, \ x \in [a_{j-1}, a_j[, \ j = 1, ..., k,$$

et $F_n(x) = 0$ pour $x < a_0$ et $x \geq a_k$.

Proposition 3.9. *Pour tout $\omega \in \Omega$, l'histogramme F_n est une densité.*

Preuve. Clairement, $F_n(x) \geq 0$ pour tout $x \in \mathbb{R}$ et

$$\int_{\mathbb{R}} f(x)dx = \frac{1}{n} \sum_{j=1}^{k} \int_{a_{j-1}}^{a_j} \frac{\mathbb{1}\{a_{j-1} \leq x_i < a_j\}}{a_j - a_{j-1}} dx = \frac{1}{n} \sum_{j=1}^{k} \mathbb{1}\{a_{j-1} \leq x_i < a_j\} = 1,$$

du fait que chaque observation x_i appartient à un et un seul intervalle $[a_{j-1}, a_j[$. □

L'histogramme est entièrement déterminé par la discrétisation $(a_j)_{j=0,...,k}$ choisie. Un choix naturel est de prendre des intervalles de tailles égales entre les valeurs extrêmes de l'échantillon. Dans ce cas, seul le nombre k de classes intervient dans la construction de l'histogramme. La qualité de l'histogramme dépend fortement du nombre de classes (cf. Figure 1).

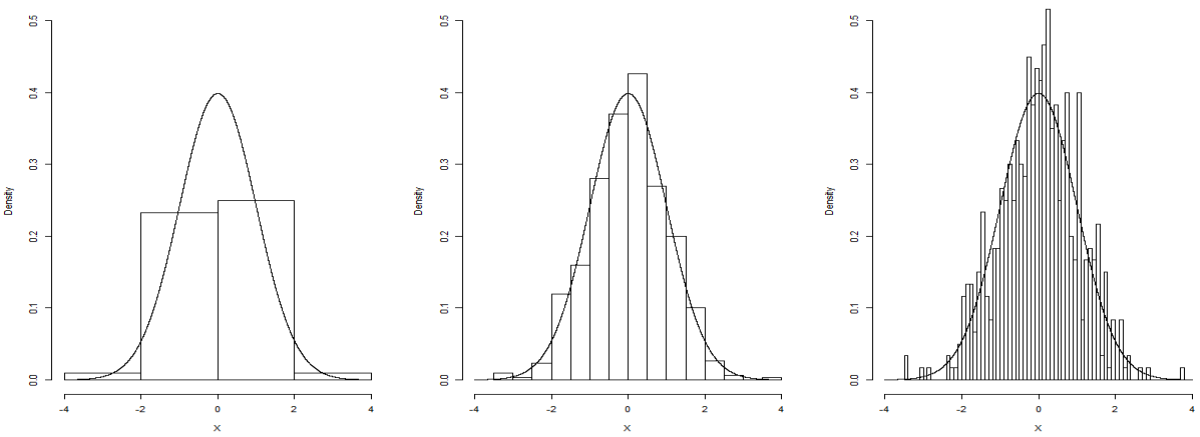

Figure 1: Histogramme obtenu avec 4 classes (à gauche), 20 classes (au centre) et 50 classes (à droite) sur un échantillon de 500 réalisations de variables aléatoires de loi normale standard.

De manière générale, le nombre optimal de classes dépend de n. En particulier, il faut que le nombre de classes $k = k_n$ vérifie $\lim_{n\to\infty} k_n = \infty$ et $\lim_{n\to\infty} k_n/n = 0$ (en pratique, on peut prendre $k_n \simeq n^{1/3}$).

3.2.2 Histogramme mobile

Pour ne pas avoir à choisir une discrétisation $(a_j)_{j=0,...,k}$, il existe une autre forme d'histogramme appelée *histogramme mobile*. L'idée est basée sur le fait que, si une fonction f est suffisamment "lisse" dans le voisinage d'un point x, alors pour des petites valeurs de $h > 0$,

$$f(x) \approx \frac{1}{2h} \int_{x-h}^{x+h} f(s)ds.$$

En particulier, il y égalité si la fonction est constante sur $]x-h, x+h]$. Dans la situation qui nous intéresse, cela revient à dire que si X a pour densité f, alors

$$f(x) \approx \frac{\mathbb{P}(x-h < X \leq x+h)}{2h}.$$

Le terme de droite est facile à estimer à partir de la fonction de répartition empirique, un estimateur naturel de $f(x)$ est donc donné par

$$f_n(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{1}\{X_i \in]x-h, x+h]\}.$$

L'estimateur f_n est appelé histogramme mobile car il est constant par morceaux mais ses classes dépendent continuement de x. L'histogramme mobile ne dépend que d'un paramètre à calibrer: h (la fenêtre). Le choix de la fenêtre est crucial et l'estimateur y est très sensible, comme on peut le voir sur le graphique suivant.

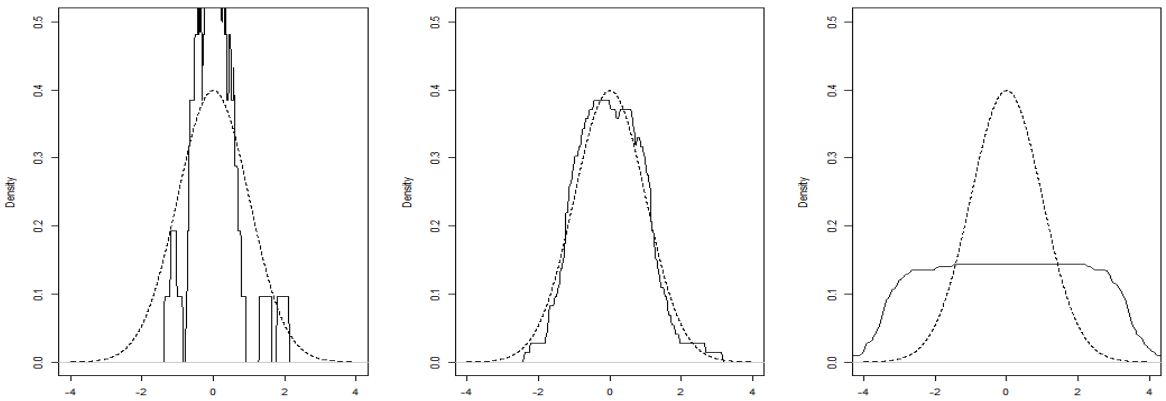

Figure 2: Histogramme mobile pour un échantillon de loi normale standard de 30 observations obtenu avec une fenêtre $h = 0.1$ (à gauche), $h = 0.7$ (au centre) et $h = 2$ (à droite) (la vraie densité est représentée en pointillée)

Proposition 3.10. *Pour tout $\omega \in \Omega$, l'histogramme mobile f_n est une densité.*

Preuve. Il est clair que $f_n \geq 0$. En utilisant que $\mathbb{1}\{x_i \in]x-h, x+h]\} = \mathbb{1}\{x_i - h \leq x < x_i + h\}$, on obtient

$$\int_{\mathbb{R}} f(x)dx = \frac{1}{2nh} \sum_{i=1}^{n} \int_{x_i-h}^{x_i+h} dx = \frac{1}{2nh} \sum_{i=1}^{n} 2h = 1. \qquad \square$$

On remarque que l'estimateur $f_n(x)$ est (en général) biaisé mais que le biais est faible pour des petites valeurs de h,

$$\mathbb{E}(f_n(x)) = \frac{F(x+h) - F(x-h)}{2h} \underset{h \to 0}{\approx} F'(x) = f(x).$$

En revanche, une petite valeur de h va entraîner une forte variance

$$\mathrm{var}(f_n(x)) = \frac{F(x+h) - F(x-h)}{2h} \; \frac{1 - F(x+h) + F(x-h)}{2nh} \underset{h \to 0}{\approx} \frac{f(x)}{2nh}.$$

L'estimateur $f_n(x)$ est donc optimal pour une fenêtre h ni trop grande, ni trop faible, l'objectif étant de trouver un compromis entre le biais et la variance. En particulier, pour que F_n converge simplement dans $\mathbb{L}^2$ vers f, il faut choisir $h = h_n$ tel que $h_n \to 0$ et $nh_n \to \infty$ de manière à faire tendre simultanément le biais et la variance vers zéro.

Théorème 3.11. *Si f est de classe $\mathcal{C}^2$ et de dérivée seconde f'' bornée en valeur absolue par $L > 0$, alors l'histogramme mobile f_n vérifie*

$$\mathbb{E}\Big[\big(f_n(x) - f(x)\big)^2\Big] \leq \frac{h^4 L^2}{36} + \frac{f(x)}{2nh} + \frac{hL}{12n}.$$

Preuve. On fait la décomposition suivante (EQM = biais^2 + variance)

$$\mathbb{E}\Big[\big(f_n(x) - f(x)\big)^2\Big] = \big[\mathbb{E}\big(f_n(x) - f(x)\big)\big]^2 + \operatorname{var}\big(f_n(x)\big).$$

Pour majorer le carré du biais on écrit un développement de Taylor-Lagrange à l'ordre 3: il existe $\xi_1 \in]x-h, x[$ et $\xi_2 \in]x, x+h[$ tels que

$$\begin{aligned} F(x+h) &= F(x) + hf(x) + \frac{h^2}{2} f'(x) + \frac{h^3}{6} f''(\xi_1) \\ F(x-h) &= F(x) - hf(x) + \frac{h^2}{2} f'(x) - \frac{h^3}{6} f''(\xi_2) \end{aligned}$$

On déduit $F(x+h) - F(x-h) = 2hf(x) + h^3(f''(\xi_1) + f''(\xi_2))/6$ et

$$\big[\mathbb{E}\big(f_n(x) - f(x)\big)\big]^2 = \left[\frac{F(x+h) - F(x-h)}{2h} - f(x)\right]^2 = \frac{h^4}{144}\big(f''(\xi_1) + f''(\xi_2)\big)^2 \leq \frac{h^4 L^2}{36}.$$

De même, comme F est croissante, $1 - F(x+h) + F(x-h) \leq 1$ et

$$\operatorname{var}\big(f_n(x)\big) = \frac{F(x+h) - F(x-h)}{2h} \; \frac{1 - F(x+h) + F(x-h)}{2nh} \leq \frac{f(x) + h^2 L/6}{2nh}. \qquad \square$$

Cette majoration de l'erreur quadratique dépend de $f(x)$ et L qui sont en général inconnus. Pour obtenir un ordre de grandeur "optimal" pour la fenêtre, on choisit $h = h_n$ tel que les deux termes dominant $h^4 L^2/36$ et $f(x)/2nh$ (lorsque $f(x) > 0$) soient du même ordre.

Corollaire 3.12. *Sous les hypothèses du théorème 3.11, si $h = h_n = Cn^{-1/5}$ pour $C > 0$, alors*

$$\mathbb{E}\Big[\big(f_n(x) - f(x)\big)^2\Big] = O(n^{-4/5}).$$

Ce résultat montre que l'histogramme mobile construit avec une fenêtre h_n de l'ordre de $n^{-1/5}$ converge simplement vers f dans $\mathbb{L}^2$, à une vitesse $n^{2/5}$ (la norme $\mathbb{L}^2$ est la racine carré de l'erreur quadratique). Cette vitesse, plus lente que la vitesse classique $\sqrt{n}$ que l'on retrouve en estimation paramétrique (par exemple pour le TCL), montre d'une certaine façon que la densité est plus difficile à estimer.

3.2.3 Estimateur à noyaux

L'estimation de densité par noyaux (parfois appelée méthode de Parzen-Rosenblatt) est une généralisation de l'histogramme mobile. L'une des motivations principales est de construire un estimateur lisse de la densité en remplaçant les indicatrices utilisées dans la construction de l'histogramme mobile par des fonctions continues.

Définition 3.13. *Un noyau $K : \mathbb{R} \to \mathbb{R}$ est une fonction paire, positive et qui vérifie*

$$\int_{-\infty}^{+\infty} K(s)ds = 1.$$

L'estimateur à noyau de f est défini par

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \ x \in \mathbb{R}.$$

Il dépend donc du choix du noyau K et de la fenêtre h. Les exemples les plus utilisés de noyaux sont

- Le noyau uniforme correspondant à l'histogramme mobile: $K(s) = \frac{1}{2}\mathbb{1}\{-1 \leq s \leq 1\}$
- Le noyau Gaussien: $K(s) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{s^2}{2})$
- Le noyau triangulaire: $K(s) = (1 - |s|)\mathbb{1}\{-1 \leq s \leq 1\}$
- Le noyau d'Epanechnicov: $K(s) = \frac{3}{4}(1 - s^2)\mathbb{1}\{-1 \leq s \leq 1\}$

Théorème 3.14. *Soit K un noyau vérifiant*

$$\int_{\mathbb{R}} s^2 K(s)ds = C_1 < \infty \quad , \quad \int_{\mathbb{R}} K^2(s)ds = C_2 < \infty \quad \text{et} \quad \int_{\mathbb{R}} s^2 K^2(s)ds = C_3 < \infty.$$

Sous les hypothèses du théorème 3.11, f_n vérifie

$$\mathbb{E}\Big[\big(f_n(x) - f(x)\big)^2\Big] \leq \frac{h^4 L^2 C_1^2}{4} + \frac{f(x)C_2}{nh} + \frac{hLC_3}{2n}.$$

Preuve. On utilise la décomposition biais-variance de l'erreur quadratique

$$\mathbb{E}\Big[\big(f_n(x) - f(x)\big)^2\Big] = \big[\mathbb{E}\big(f_n(x) - f(x)\big)\big]^2 + \operatorname{var}\big(f_n(x)\big).$$

Pour le biais, on écrit

$$\mathbb{E}\big(f_n(x)\big) = \frac{1}{h}\int_{\mathbb{R}} K\big(\frac{s-x}{h}\big) f(s)ds = \int_{\mathbb{R}} K(u)f(x+hu)du.$$

D'après la formule de Taylor à l'ordre 1, on sait qu'il existe $\xi(u) \in]x, x+hu[$ (ou $]x+hu, x[$ si $u < 0$) continu en u tel que $f(x+hu) = f(x) + huf'(x) + h^2u^2f''(\xi(u))/2$. On déduit

$$\begin{aligned}
\mathbb{E}\big(f_n(x)\big) &= \int_{\mathbb{R}} K(u)f(x)du + \int_{\mathbb{R}} K(u)huf'(x)du + \int_{\mathbb{R}} K(u)\frac{h^2u^2}{2}f''(\xi(u))du \\
&= f(x) + \frac{h^2}{2}\int_{\mathbb{R}} K(u)u^2 f''(\xi(u))du.
\end{aligned}$$

D'où $\big[\mathbb{E}\big(f_n(x) - f(x)\big)\big]^2 \leq \dfrac{h^4L^2C_1^2}{4}$. Pour la variance,

$$\operatorname{var}\big(f_n(x)\big) = \frac{1}{nh^2}\operatorname{var}\left(K\big(\frac{X-x}{h}\big)\right) = \frac{1}{nh^2}\int_{\mathbb{R}} K^2\big(\frac{s-x}{h}\big)f(s)ds - \big[\mathbb{E}\big(f_n(x)\big)\big]^2$$

avec

$$\begin{aligned}\frac{1}{h}\int_{\mathbb{R}} K^2(\frac{s-x}{h})f(s)ds &= \int_{\mathbb{R}} K^2(u)f(x+hu)du \\ &= f(x)\int_{\mathbb{R}} K^2(u)du + f'(x)h\int_{\mathbb{R}} uK^2(u)du + \frac{h^2}{2}\int_{\mathbb{R}} u^2K^2(u)f''(\xi(u))du.\end{aligned}$$

On sait par l'inégalité de Cauchy-Schwarz que

$$\left(\int_{\mathbb{R}} |u|K^2(u)du\right)^2 \leq \int_{\mathbb{R}} u^2K^2(u)du \int_{\mathbb{R}} K^2(u)du = C_3C_2 < \infty.$$

Comme K est paire, K^2 l'est aussi et $\int_{\mathbb{R}} uK^2(u)du = 0$. On obtient donc la majoration de la variance

$$\operatorname{var}\big(f_n(x)\big) \leq \frac{f(x)C_2}{nh} + \frac{hLC_3}{2n}. \qquad \square$$

3.3 Quantiles

3.3.1 Fonction de quantile

Définition 3.15. *Soit $\alpha \in]0,1[$, on appelle quantile d'ordre α d'une variable aléatoire X, tout réel q_α vérifiant*

$$\mathbb{P}(X \leq q_\alpha) \geq \alpha \quad \textit{et} \quad \mathbb{P}(X \geq q_\alpha) \geq 1-\alpha.$$

Si X admet une densité f, alors on sait que $]0,1[\subseteq F(\mathbb{R})$, ce qui implique qu'il existe pour tout $\alpha \in]0,1[$, un quantile q_α tel que

$$\mathbb{P}(X \leq q_\alpha) = \alpha \quad \text{et} \quad \mathbb{P}(X \geq q_\alpha) = 1-\alpha.$$

Si de plus le support de f est connexe, alors F est une bijection et le quantile d'ordre α est unique pour tout $\alpha \in]0,1[$.

Remarque 3.16. *La médiane est le quantile d'ordre* 0.5. *Les premier et troisième quartiles sont les quantiles d'ordre* 0.25 *et* 0.75 *respectivement.*

Définition 3.17. *La fonction quantile F^- est l'inverse généralisée de F définie par*

$$F^-(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \ u \in]0,1[.$$

La fonction quantile F^- est toujours définie, même lorsque F n'est pas continue ou est constante sur un intervalle. Le graphe de F^- est le symétrique par rapport à la première bissectrice du graphe de F pour lequel une partie constante devient un point de discontinuité.

Remarque 3.18. *Par continuité à droite de F, on a que $F \circ F^-(u) \geq u$ pour tout $u \in]0,1[$. Lorsque F est continue, $F^-(\alpha)$ est simplement le plus petit quantile d'ordre α. Si de plus, F est une bijection, la fonction quantile est la fonction réciproque F^{-1} de F.*

Lemme 3.19. *Soit U une variable aléatoire de loi uniforme sur $]0,1[$, $F^-(U)$ est une variable aléatoire de fonction de répartition F.*

Preuve. Soit $(x_n)_{n\in\mathbb{N}}$ une suite décroissante de $\{x \in \mathbb{R} : F(x) \geq u\}$ qui converge vers sa borne inférieure $F^-(u)$. Par construction, $F(x_n) \geq u$ et par la continuité à droite de F,

$$F \circ F^-(u) = \lim_{n\to\infty} F(x_n) \geq u.$$

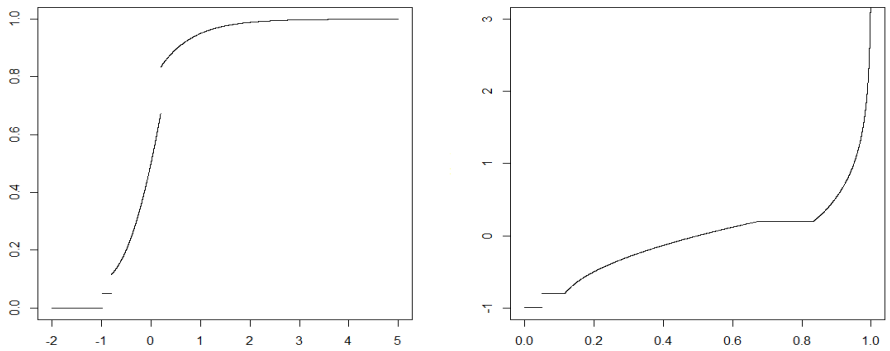

Figure 3: Fonction de répartition (à gauche) et son inverse généralisée (à droite)

Ainsi, pour tout $x \in \mathbb{R}$,

$$\mathbb{P}(F^-(U) \leq x) \leq \mathbb{P}\big(F \circ F^-(U) \leq F(x)\big) \leq \mathbb{P}\big(U \leq F(x)\big) = F(x).$$

Soit $v > u$, le fait que F est croissante entraîne $\{x \in \mathbb{R} : F(x) \geq v\} \subseteq \{x \in \mathbb{R} : F(x) \geq u\}$ et

$$F^-(v) = \inf\{x \in \mathbb{R} : F(x) \geq v\} \geq \inf\{x \in \mathbb{R} : F(x) \geq u\} = F^-(u),$$

F^- est donc croissante. Comme $F^- \circ F(x) = \inf\{y \in \mathbb{R} : F(y) \geq F(x)\} \leq x$, on conclut

$$F(x) = \mathbb{P}\big(U \leq F(x)\big) \leq \mathbb{P}\big(F^-(U) \leq F^- \circ F(x)\big) \leq \mathbb{P}\big(F^-(U) \leq x\big). \qquad \square$$

Remarque 3.20. *Contrairement au lemme 3.6, ce résultat ne nécessite pas d'hypothèse de continuité et est valable pour toute fonction de répartition F.*

3.3.2 Quantiles empiriques

Pour l'estimation des quantiles à partir d'un échantillon iid $X_1, ..., X_n$, on définit la fonction de quantile empirique

$$F_n^-(u) = \inf\{x \in \mathbb{R} : F_n(x) \geq u\}.$$

C'est la fonction de quantile associée à la fonction de répartition empirique. Le quantile empirique d'ordre α peut se définir également à partir des statistiques d'ordre de l'échantillon $X_1, ..., X_n$.

Définition 3.21. *Les statistiques d'ordre $X_{(1)}, ..., X_{(n)}$ sont les valeurs ordonnées dans l'ordre croissant de l'échantillon $X_1, ..., X_n$.*

Les statistiques d'ordre $X_{(k)}$ pour $k = 1, ..., n$ sont des variables aléatoires qui vérifient par construction $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$. Pour le résultat suivant, on définit la fonction $x \mapsto \lceil x \rceil$ qui à $x \in \mathbb{R}$ associe le plus petit entier supérieur ou égal à x.

Proposition 3.22. *Pour tout $\alpha \in]0, 1[$, $F_n^-(\alpha) = X_{(\lceil n\alpha \rceil)}$.*

Preuve. F_n est constante par morceaux et vérifie $F_n(x) = 0$ si $x < X_{(1)}$, $F_n(x) = 1$ si $x \geq X_{(n)}$ et

$$F_n(x) = \frac{k}{n}\ ,\ x \in [X_{(k)}, X_{(k+1)}[\ ,\ k = 1, ..., n-1.$$

On a en particulier $F_n(X_{(\lceil n\alpha \rceil)}) = \dfrac{\lceil n\alpha \rceil}{n} \geq \alpha$ et $\forall x < X_{(\lceil n\alpha \rceil)}$, $F_n(x) \leq \dfrac{\lceil n\alpha \rceil - 1}{n} < \alpha$. Ainsi,

$$F_n^-(\alpha) = \inf\{x \in \mathbb{R} : F_n(x) \geq \alpha\} = X_{(\lceil n\alpha \rceil)}. \qquad \square$$

Remarque 3.23. *Le quantile empirique est un quantile de l'ensemble $\{x_1, ..., x_n\}$. Par exemple, la médiane empirique $F_n^-(0.5)$ vaut $X_{(\frac{n}{2}+1)}$ lorsque n est pair et $X_{(\frac{n+1}{2})}$ lorsque n est impair.*

Pour étudier le comportement des quantiles empiriques, on étudie les statistiques d'ordre.

Théorème 3.24. *La k-ième statistique d'ordre $X_{(k)}$ a pour densité*

$$f_{(k)}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1 - F(x)\big)^{n-k} f(x),$$

pour tout x point de continuité de f.

Preuve. Soit $h > 0$, on a pour tout $x \in \mathbb{R}$

$$\begin{aligned}
\mathbb{P}(x < X_{(k)} \leq x + h) &= \mathbb{P}\Big(\#\{i : X_i \leq x\} \leq k-1,\ \#\{i : X_i \leq x+h\} \geq k\Big) \\
&= \sum_{j=0}^{k-1} \sum_{\ell=k-j}^{n-j} \mathbb{P}\Big(\#\{i : X_i \leq x\} = j,\ \#\{i : x < X_i \leq x+h\} = \ell\Big) \\
&= \sum_{j=0}^{k-1} \sum_{\ell=k-j}^{n-j} \binom{n}{j}\binom{n-j}{\ell} F(x)^j \big(F(x+h) - F(x)\big)^\ell \big(1 - F(x)\big)^{n-j-\ell}.
\end{aligned}$$

On déduit

$$f_{(k)}(x) = \lim_{h \to 0} \frac{1}{h}\ \mathbb{P}(x < X_{(k)} \leq x + h) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}\big(1 - F(x)\big)^{n-k} f(x),$$

les termes pour $\ell \geq 2$ étant négligeables quand $h \to 0$. □

La statistique d'ordre $X_{(\lceil n\alpha \rceil)} = F_n^-(\alpha)$ est un estimateur du quantile $q_\alpha = F^-(\alpha)$ qui est convergent si F est bijective sur un voisinage de q_α.

Proposition 3.25. *Si F est de classe $\mathcal{C}^1$ dans un voisinage de q_α et $F'(q_\alpha) = f(q_\alpha) > 0$, alors $X_{(\lceil n\alpha \rceil)}$ converge en probabilité vers q_α quand $n \to \infty$ et*

$$\sqrt{n}(X_{(\lceil n\alpha \rceil)} - q_\alpha) = O_P(1).$$

Preuve. Soit $U_i = F(X_i), i = 1, ..., n$, comme F est croissante, $U_{(i)} = F(X_{(i)})$ presque sûrement. Montrons que $U_{(\lceil n\alpha \rceil)}$ converge dans $\mathbb{L}^2$ vers α. D'après le lemme 3.6, on sait que $U_1, ..., U_n$ est un échantillon iid de loi uniforme sur $[0, 1]$. De plus, par le théorème 3.24, $U_{(\lceil n\alpha \rceil)}$ suit une loi beta $B(\lceil n\alpha \rceil, n - \lceil n\alpha \rceil + 1)$ où on rappelle que la densité de la loi $B(a, b)$ est donnée par

$$\beta_{a,b}(u) = \frac{(a+b-1)!}{(a-1)!(b-1)!} u^{a-1}(1-u)^{b-1}\ ,\ u \in]0, 1[,$$

pour $a, b \in \mathbb{N}$. En particulier, l'espérance et la variance de la loi $B(a, b)$ sont

$$\mathbb{E}(B(a, b)) = \frac{a}{a+b} \quad \text{et} \quad \text{var}(B(a, b)) = \frac{ab}{(a+b)^2(a+b+1)}.$$

On a donc

$$\mathbb{E}\big(U_{(\lceil n\alpha \rceil)}\big) = \frac{\lceil n\alpha \rceil}{n+1} \quad \text{et} \quad \text{var}\left(U_{(\lceil n\alpha \rceil)}\right) = \frac{\lceil n\alpha \rceil(n - \lceil n\alpha \rceil + 1)}{(n+1)^2(n+2)} = \frac{\lceil n\alpha \rceil}{(n+1)(n+2)} - \frac{\lceil n\alpha \rceil^2}{(n+1)^2(n+2)}.$$

On en déduit l'erreur quadratique moyenne

$$\mathbb{E}\big(U_{(\lceil n\alpha\rceil)}-\alpha\big)^2=\left(\frac{\lceil n\alpha\rceil}{n+1}-\alpha\right)^2+\frac{\lceil n\alpha\rceil(n-\lceil n\alpha\rceil+1)}{(n+1)^2(n+2)}.$$

On remarque que

$$\left(\frac{\lceil n\alpha\rceil}{n+1}-\alpha\right)^2=\left(\frac{\lceil n\alpha\rceil-n\alpha-\alpha}{n+1}\right)^2\leq\frac{1}{(n+1)^2}$$

et

$$\frac{\lceil n\alpha\rceil}{(n+1)(n+2)}-\frac{\lceil n\alpha\rceil^2}{(n+1)^2(n+2)}=\frac{\alpha(1-\alpha)}{n+2}+o\left(\frac{1}{n}\right).$$

On en déduit que $n\mathbb{E}\big(U_{(\lceil n\alpha\rceil)}-\alpha\big)^2=O(1)$ et donc en particulier

$$\sqrt{n}\big(U_{(\lceil n\alpha\rceil)}-\alpha\big)=O_P(1).$$

Par hypothèse, F est bijective dans un voisinage de q_α, on a donc que $U_{(\lceil n\alpha\rceil)}=F(X_{(\lceil n\alpha\rceil)})$ converge vers α si, et seulement si, $X_{(\lceil n\alpha\rceil)}$ converge vers $F^-(\alpha)=q_\alpha$. Par continuité de F^- au voisinage de α,

$$\sqrt{n}\big(F^-(U_{(\lceil n\alpha\rceil)})-F^-(\alpha)\big)\overset{ps}{=}\sqrt{n}\big(X_{(\lceil n\alpha\rceil)}-q_\alpha\big)=O_P(1).\qquad\square$$

Remarque 3.26. *Sous les hypothèses du théorème, on peut montrer un résultat plus fort, à savoir la normalité asymptotique du quantile empirique* $X_{(\lceil n\alpha\rceil)}$,

$$\sqrt{n}\big(X_{(\lceil n\alpha\rceil)}-q_\alpha\big)\xrightarrow[n\to\infty]{loi}\mathcal{N}\left(0,\frac{\alpha^2(1-\alpha)^2}{f(q_\alpha)^2}\right).$$

4 Analyse bivariée

L'analyse bivariée consiste à étudier les liens qui peuvent exister entre deux variables aléatoires. Dans ce chapitre, les observations sont des réalisations indépendantes $(X_1,Y_1),...,(X_n,Y_n)$ d'un couple de variables aléatoires (X,Y) de loi μ sur $\mathbb{R}^2$ et de lois marginales μ_X et μ_Y. On supposera que X et Y sont des variables non constantes et de variance finie.

4.1 Corrélation

On rappelle que deux variables aléatoires X et Y de carré intégrable sont dites corrélées si leur covariance $\mathrm{cov}(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$ est non-nulle. La corrélation entre X et Y est alors mesurée par

$$\mathrm{cor}(X,Y):=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}},$$

et $\mathrm{cor}(X,Y)=0$ si une des variables est constante presque sûrement. La corrélation vérifie les propriétés suivantes:

- Par l'inégalité de Cauchy-Schwarz, $-1\leq\mathrm{cor}(X,Y)\leq 1$.
- $\mathrm{cor}(X,Y)=1$ si, et seulement si, $Y=aX+b$ pour un $a>0$ et $b\in\mathbb{R}$.
- $\mathrm{cor}(X,Y)=-1$ si, et seulement si, $Y=aX+b$ pour un $a<0$ et $b\in\mathbb{R}$.

On suppose maintenant que l'on cherche à évaluer s'il y a un lien affine entre X et et Y à partir d'un échantillon iid $(X_1, Y_1), ..., (X_n, Y_n)$ de même loi que (X, Y). On note $\overline{X}_n$ et $\overline{Y}_n$ les moyennes empiriques de X et Y calculées sur l'échantillon,

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

qui permettent d'estimer les moyennes théoriques $\mathbb{E}(X)$ et $\mathbb{E}(Y)$. De la même façon, on définit les variances et covariance empiriques,

$$\hat{\sigma}^2(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \ , \ \hat{\sigma}^2(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \ , \ \hat{\sigma}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$$

Définition 4.1. *Le coefficient de corrélation de Pearson est donné par*

$$\hat{\rho} = \frac{\hat{\sigma}(X,Y)}{\hat{\sigma}(X)\ \hat{\sigma}(Y)}.$$

Le coefficient de Pearson mesure la corrélation entre les vecteurs $(X_1, ..., X_n)$ et $(Y_1, ..., Y_n)$ dans $\mathbb{R}^n$. Il permet d'estimer la corrélation théorique $\mathrm{cor}(X, Y)$ et comme elle, est compris entre -1 et 1. Le lien affine est d'autant plus net que $\hat{\rho}$ est proche de 1 ou -1 (1 pour une relation croissante et -1 pour une relation décroissante).

Remarque 4.2. *Un coefficient de corrélation de Pearson proche de zéro ne signifie pas que les variables sont indépendantes mais seulement qu'une modélisation de Y pas une fonction affine de X n'est pas pertinente. Par exemple, une relation du type $Y_i = X_i^2$ pour des X_i distribués sur $[-1, 1]$ donnera une valeur de $\hat{\rho}$ proche de zéro.*

Pour mettre en évidence l'existence d'une relation monotone (pas forcément linéaire) entre X et Y du type $Y_i = \phi(X_i) + \epsilon_i$ où ϕ est une fonction monotone, on peut utiliser le *coefficient de corrélation de Spearman* qui se construit à partir des rangs des variables X et Y dans l'échantillon. En notant $R_i \in \{1, ..., n\}$ le rang de X_i (on a donc $X_{R_1} \geq ... \geq X_{R_n}$, en privilégiant en cas d'égalité l'indice le plus petit) et S_i le rang de Y_i, le coefficient de corrélation de Spearman est défini par

$$\hat{\gamma} = \frac{\hat{\sigma}(R,S)}{\hat{\sigma}(R)\ \hat{\sigma}(S)}.$$

L'idée intuitive du coefficient de corrélation de Spearman est de dire que si Y est une fonction monotone de X alors les rangs $R_i = S_i$ dans le cas d'une relation croissante, et $R_i = n - S_i$ dans le cas d'une relation décroissante. Dans les deux cas, les variables R_i et S_i sont liées linéairement.

En conséquence immédiate de la loi forte des grands nombres, $\hat{\rho}$ converge presque sûrement vers $\mathrm{cor}(X, Y)$ quand $n \to \infty$. On a de plus le comportement asymptotique suivant lorsque X et Y sont indépendantes.

Théorème 4.3. *Si X et Y sont indépendantes, alors*

$$\sqrt{n}\ \hat{\rho} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

Preuve. On pose $Z_i = (X_iY_i, X_i, Y_i)^\top$, $i = 1, ..., n$ et

$$\overline{Z}_n = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n X_iY_i \\ \frac{1}{n}\sum_{i=1}^n X_i \\ \frac{1}{n}\sum_{i=1}^n Y_i \end{bmatrix} \quad \text{et} \quad \Sigma = \mathrm{var}(Z_1) = \begin{bmatrix} \mathrm{var}(XY) & \mathrm{cov}(XY,X) & \mathrm{cov}(XY,Y) \\ \mathrm{cov}(XY,X) & \mathrm{var}(X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(XY,Y) & \mathrm{cov}(X,Y) & \mathrm{var}(Y) \end{bmatrix}.$$

Par le théorème central limite vectoriel, on sait que

$$\sqrt{n}\big(\overline{Z} - \mathbb{E}(Z)\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Sigma).$$

Soit $\mathsf{c}_n = \big(1, -\overline{Y}_n, -\mathbb{E}(X)\big)$, on sait par le théorème de Slutsky que

$$\sqrt{n}\ \mathsf{c}_n\big(\overline{Z}_n - \mathbb{E}(Z)\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \mathsf{c}\Sigma\mathsf{c}^\top)$$

où $\mathsf{c} = (1, -\mathbb{E}(Y), -\mathbb{E}(X))$. Il reste à remarquer que si X et Y sont indépendantes, alors

$$\hat{\sigma}(X,Y) = \mathsf{c}_n\big(\overline{Z}_n - \mathbb{E}(Z)\big) \quad \text{et} \quad \mathsf{c}\Sigma\mathsf{c}^\top = \mathrm{var}(X)\,\mathrm{var}(Y).$$

On déduit le résultat final en utilisant une nouvelle fois le théorème de Slutsky. □

4.2 Modélisation non-paramétrique

Pour aller plus loin dans l'analyse bivariée, on peut rechercher la meilleure approximation de Y par une fonction de X, sans se limiter aux fonctions affines. Rappelons la définition suivante.

Définition 4.4. *Soit Y une variable aléatoire réelle intégrable, l'espérance conditionnelle de Y sachant X, notée $\mathbb{E}(Y|X)$, est une variable aléatoire fonction de X qui vérifie*

$$\forall g \text{ continue bornée }, \mathbb{E}\big(Yg(X)\big) = \mathbb{E}\big(\mathbb{E}(Y|X)g(X)\big).$$

On rappelle que que l'espérance conditionnelle existe et est unique presque sûrement. Toute fonction ϕ^* vérifiant $\mathbb{E}(Y|X) \overset{ps}{=} \phi^*(X)$, appelée fonction de regression, est par conséquent unique μ_X-presque partout.

Théorème 4.5. *Si Y est de carré intégrable, alors ϕ^* minimise*

$$\phi \mapsto \mathbb{E}\big(Y - \phi(X)\big)^2$$

parmi les fonctions ϕ mesurables.

Remarque 4.6. *Dans le cas où $\mathbb{E}(Y^2) < \infty$, l'espérance conditionnelle $\mathbb{E}(Y|X)$ s'interprète comme la projection orthogonale de Y sur le sous-espace vectoriel des variables aléatoires de carré intégrable mesurables par rapport à la tribu engendrée par X.*

L'espérance conditionnelle vérifie les propriétés suivantes.

- $\mathbb{E}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(Y)$.
- Si X et Y sont indépendantes, alors $\mathbb{E}(Y|X) = \mathbb{E}(Y)$.
- Si Y est une fonction de X, alors $\mathbb{E}(Y|X) = Y$.

Théorème 4.7. *Si X, Y et (X,Y) ont pour densités respectives f_X, f_Y et f_{XY}, alors pour tout x tel que $f_X(x) > 0$,*

$$\phi^*(x) = \int_{\mathbb{R}} y\ \frac{f_{XY}(x,y)}{f_X(x)} dy.$$

Preuve dans le cas $\mathbb{E}(Y^2) < \infty$. Montrons que la fonction ϕ^* définie dans le théorème minimise $\phi \mapsto \mathbb{E}\big(Y - \phi(X)\big)^2$. On écrit $\mathbb{E}\big(Y - \phi(X)\big)^2 = \mathbb{E}(Y^2) + \mathbb{E}\big(\phi(X)^2\big) - 2\mathbb{E}\big(Y\phi(X)\big)$ avec par le théorème de Fubini

$$\begin{aligned}\mathbb{E}\big(\phi(X)^2\big) - 2\mathbb{E}\big(Y\phi(X)\big) &= \int_{\mathbb{R}} \phi(x)^2 \; f_X(x)dx - 2\int_{\mathbb{R}^2} \phi(x) y f_{XY}(x,y)dxdy \\ &= \int_{\mathbb{R}} \Big(\phi(x)^2 - 2\phi(x)\frac{\int_{\mathbb{R}} y \; f_{XY}(x,y)dy}{f_X(x)}\Big) f_X(x)dx.\end{aligned}$$

La quantité entre paranthèses est minimale en $\phi(x) = \int_{\mathbb{R}} y \; f_{XY}(x,y)dy / f_X(x)$, pour tout $x \in \mathbb{R}$, ce qui termine la preuve. □

Preuve dans le cas général. Montrons que la fonction ϕ^* définie dans le théorème vérifie, pour toute fonction g continue bornée, $\mathbb{E}\big(Yg(X)\big) = \mathbb{E}\big(\phi^*(X)g(X)\big)$. On a

$$\mathbb{E}\big(\phi^*(X)g(X)\big) = \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} y \; \frac{f_{XY}(x,y)}{f_X(x)} dy\Big) g(x) f_X(x)dx = \int_{\mathbb{R}^2} y g(x) f_{XY}(x,y)dxdy = \mathbb{E}\big(Yg(X)\big),$$

par le théorème de Fubini. □

Lorsque les variables sont à densités, l'espérance conditionnelle $\phi^*(x) = \mathbb{E}(Y|X = x)$ est simplement l'espérance de Y sous la densité conditionnelle de Y sachant $X = x$, donnée par la formule de Bayes

$$f_{Y|X}(y,x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

Il existe de nombreuses méthodes pour estimer la fonction de régression ϕ^* dans le modèle d'échantillonage. Une des plus couramment utilisée est la méthode de Nadaraya-Watson, qui est basée sur une estimation à noyaux. Soit K un noyau, on définit l'estimateur

$$\hat{\phi}(x) = \frac{1}{\sum_{i=1}^n K\big(\frac{X_i - x}{h}\big)} \sum_{i=1}^n Y_i \; K\Big(\frac{X_i - x}{h}\Big), \; x \in \mathbb{R}$$

où $h > 0$ est la fenêtre à calibrer. L'idée de l'estimateur de Nadaraya-Watson est d'estimer l'espérance de Y sachant $X = x$ par une moyenne des Y_i, pondérée en fonction de la proximité des X_i avec x. En effet, on peut écrire $\hat{\phi}(x) = \sum_{i=1}^n W_i(x) Y_i$ où

$$W_i(x) = \frac{K\big(\frac{X_i - x}{h}\big)}{\sum_{i=1}^n K\big(\frac{X_i - x}{h}\big)}$$

est un poids qui mesure la proximité entre X_i et x, renormalisé pour avoir $\sum_{i=1}^n W_i(x) = 1$.

Comme pour l'estimation à noyaux d'une densité, on peut montrer sous des conditions de régularités de f_{XY} qu'en prenant $h = h_n$ tel que $\lim_{n\to\infty} h_n = 0$ et $\lim_{n\to\infty} nh_n = +\infty$, alors $\hat{\phi}(x)$ converge en probabilité vers $\phi^*(x)$. En effet, on sait dans un premier temps que sous ces conditions sur h_n, le théorème 3.14 entraîne

$$\frac{1}{nh_n} \sum_{i=1}^n K\Big(\frac{X_i - x}{h_n}\Big) \xrightarrow[n\to\infty]{\mathbb{P}} f_X(x).$$

En utilisant la propriété $\mathbb{E}\big(Yg(X)\big) = \mathbb{E}\big(\phi^*(X)g(X)\big)$, on obtient également

$$\frac{1}{nh_n} \sum_{i=1}^n Y_i \; K\Big(\frac{X_i - x}{h_n}\Big) \xrightarrow[n\to\infty]{\mathbb{P}} \phi^*(x) f_X(x),$$

ce qui permet de montrer la convergence de l'estimateur de Nadaraya-Watson.

5 Intervalles de confiance

Jusqu'à présent, nous avons vu des méthodes d'estimation qui ont pour objectif, étnat donné un paramètre inconnu, de proposer une valeur en fonction des observations. Plutôt que de se limiter à une valeur, un point de vue différent consiste à proposer un ensemble de valeurs probables.

Définition 5.1. *On appelle intervalle de confiance pour θ de niveau $\alpha \in [0,1]$, un intervalle aléatoire $I = I(X_1, ..., X_n)$ construit à partir des observations, tel que*

$$\mathbb{P}(\theta \in I) = 1 - \alpha.$$

La construction d'un intervalle de confiance dépend du modèle et il est difficile de donner une recette générale. Dans ce chapitre, nous verrons les exemples les plus classiques.

5.1 IC pour la moyenne dans un modèle Gaussien

On suppose ici que les observations sont iid de loi normale $\mathcal{N}(\mathsf{m}, \sigma^2)$. On s'intéresse à la construction d'un intervalle de confiance pour la moyenne m que l'on suppose inconnue. La construction se fait différemment selon si la variance est connue ou inconnue.

1. *Variance connue.* On suppose que les observations $X_1, ..., X_n$ sont distribuées suivant une loi normale de variance σ^2 connue et de moyenne m inconnue. Dans ce modèle on sait que

$$\hat{\mathsf{m}} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\mathsf{m}, \frac{\sigma^2}{n}\right).$$

La loi de l'estimateur est connue en fonction de m. En particulier, on peut se ramener à une variable aléatoire normale centrée réduite

$$S = \sqrt{n}\,\frac{\hat{\mathsf{m}} - \mathsf{m}}{\sigma} \sim \mathcal{N}(0,1).$$

On insiste ici sur le fait que l'on n'observe pas une réalisation de S puisqu'elle dépend du paramètre inconnu m. Cependant, la moyenne m est la seule inconnue dans l'expression de S, ce qui est crucial pour permettre la construction de l'intervalle de confiance. La symétrie de la densité Gaussienne nous permet d'écrire

$$\mathbb{P}(-q_{1-\frac{\alpha}{2}} \leq S \leq q_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

où $q_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite. On cherche maintenant à isoler m dans l'expression de S de manière à en obtenir un encadrement

$$\mathbb{P}\left(-q_{1-\frac{\alpha}{2}} \leq \sqrt{n}\,\frac{\hat{\mathsf{m}} - \mathsf{m}}{\sigma} \leq q_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(\hat{\mathsf{m}} - \frac{q_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}} \leq \mathsf{m} \leq \hat{\mathsf{m}} + \frac{q_{1-\frac{\alpha}{2}}\,\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

On remarque que les bornes de l'intervalles sont fonctions des observations et donc connues. On obtient ainsi un intervalle de confiance pour m avec une marge d'erreur de α qui correspond à la probabilité que la vraie valeur du paramètre ne soit pas dans l'intervalle. Par exemple, si l'on cherche un intervalle de confiance à 95%, on prend $\alpha = 0,05$ ce qui correspond à $q_{1-\frac{\alpha}{2}} \approx 1,96$ et

$$I = \left[\hat{\mathsf{m}} - \frac{1,96\sigma}{\sqrt{n}}, \hat{\mathsf{m}} + \frac{1,96\sigma}{\sqrt{n}}\right].$$

2. *Variance inconnue.* Ici, on ne peut pas utiliser la statistique précédente car elle dépend de σ^2. On remplace donc la variance par un estimateur sans biais $\hat{\sigma}^2$. Précisément, on définit

$$\hat{\mathsf{m}} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mathsf{m}})^2 \quad \text{et} \quad S = \sqrt{n-1}\ \frac{\hat{\mathsf{m}} - \mathsf{m}}{\hat{\sigma}} \sim \mathcal{T}(n-1).$$

Il est important de remarquer que la loi de S n'est plus une loi normale mais une loi de Student à $n-1$ degrés de liberté (on admettra ce résultat qui est une conséquence du théorème de Cochran). La valeur m étant la seule inconnue dans l'expression de S, il est maintenant possible de construire un intervalle de confiance à partir des quantiles de la loi de Student.

5.2 IC pour la variance dans un modèle Gaussien

On suppose maintenant que la variance σ^2 des observations est inconnue. On cherche par le même procédé que précédemment à obtenir un encadrement de σ^2 avec probabilité $1-\alpha$ en se basant sur un estimateur $\hat{\sigma}^2$. On distingue encore deux cas, selon si la moyenne m est connue ou non.

1. *Moyenne connue.* On considère l'estimateur

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathsf{m})^2$$

et on utilise la propriété suivante

$$S = \frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n).$$

Soit $k_{\alpha,n}$ le quantile d'ordre α de la loi du Khi deux à n degrés de liberté, c'est-à-dire la valeur qui vérifie $\mathbb{P}(\chi^2(n) \geq k_{\alpha,n}) = 1-\alpha$. On a

$$\mathbb{P}(k_{\frac{\alpha}{2},n} \leq S \leq k_{1-\frac{\alpha}{2},n}) = 1-\alpha \Longleftrightarrow \mathbb{P}\left(\frac{n\hat{\sigma}^2}{k_{1-\frac{\alpha}{2},n}} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{k_{\frac{\alpha}{2},n}}\right) = 1-\alpha.$$

2. *Moyenne inconnue.* Dans ce cas, la méthode est quasiment identique, à la différence que l'on remplace la moyenne m par son estimateur empirique $\hat{\mathsf{m}}$. On pose

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mathsf{m}})^2 \quad \text{et} \quad S = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1).$$

L'estimateur $\hat{\sigma}^2$ que l'on considère ici est la version débiaisée de l'estimateur empirique. Le quantile utilisé pour la construction de l'intervalle de confiance est celui de la loi du Khi-2 à $n-1$ degrés de liberté (et non n comme dans l'exemple précédent).

5.3 IC asymptotiques

Si la taille n de l'échantillon est assez grande pour que l'approximation Gaussienne du théorème central limite soit satisfaisante, on peut généraliser les intervalles de confiance obtenus précédemment à des modèles non Gaussiens. On suppose ici que le modèle est paramétré par la moyenne θ et que la variance est fonction de la moyenne. On a donc $\mathbb{E}(X) = \theta$ et $\mathrm{var}(X) = \sigma(\theta)^2$ où $\sigma(.)$ est une fonction connue, que l'on suppose continue et strictement positive. On cherche à construire un intervalle de confiance pour θ.

Soit $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$, on remarque dans un premier temps que par continuité de $\sigma(.)$, $\sigma(\hat{\theta})$ converge en probabilité vers $\sigma(\theta) > 0$. Ainsi, par le lemme de Slutsky,

$$S = \sqrt{n}\,\frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \xrightarrow{loi} \mathcal{N}(0,1).$$

On déduit que S appartient à l'intervalle $[-1,96, 1,96]$ avec probabilité proche de 95% pour de grandes valeurs de n. Ainsi,

$$I = \left\{\theta : -1,96 \leq \sqrt{n}\frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})} \leq 1,96\right\} = \left[\hat{\theta} - \frac{1,96\ \sigma(\hat{\theta})}{\sqrt{n}}\ ,\ \hat{\theta} + \frac{1,96\ \sigma(\hat{\theta})}{\sqrt{n}}\right],$$

est un intervalle de confiance asymptotiquement de niveau 95%.

Exemples

1. Expérience de Bernoulli. On cherche un intervalle de confiance à 95% pour le paramètre p dans le modèle de Bernoulli. On sait que $\sqrt{n}(\hat{p}-p) \xrightarrow{loi} \mathcal{N}\big(0, p(1-p)\big)$. On a donc par le lemme de Slutsky,

$$\sqrt{n}\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

   Ainsi, pour n grand, $\mathbb{P}\left(-1,96 \leq \sqrt{n}(\hat{p}-p)/\sqrt{\hat{p}(1-\hat{p})} \leq 1,96\right) \approx 0,95$, ou encore

$$\mathbb{P}\left(\hat{p} - \frac{1,96\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \leq p \leq \hat{p} + \frac{1,96\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right) \approx 0,95.$$

2. Modèle de Poisson. Si les observations suivent une loi de Poisson de paramètre λ, alors moyenne et variance sont égales et

$$\sqrt{n}\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1),$$

   où $\hat{\lambda}$ est la moyenne empirique. Par le même raisonnement, on obtient un intervalle à 95%,

$$\mathbb{P}\left(\hat{\lambda} - \frac{1,96\sqrt{\hat{\lambda}}}{\sqrt{n}} \leq \lambda \leq \hat{\lambda} + \frac{1,96\sqrt{\hat{\lambda}}}{\sqrt{n}}\right) \approx 0,95.$$

6 Tests

En statistique, choisir un modèle revient à fixer des conditions sur la loi des observations. Lorsque l'on souhaite vérifier si ces conditions sont raisonnables, on peut utiliser un *test statistique*. Formellement, un test est une variable aléatoire construite à partir des observations qui prend la valeur 1 si on rejette l'hypothèse, ou 0 si on ne rejette pas l'hypothèse. La décision de rejeter ou non l'hypothèse est toujours sensible à l'aléa des observations, ce qui nous oblige à fixer à l'avance une marge d'erreur α qui correspond à la probabilité de rejeter l'hypothèse alors qu'elle est vraie. Pour tester une hypothèse $\mathcal{H}_0$, on construit une variable aléatoire dont la loi est connue si $\mathcal{H}_0$ est vraie. Si la valeur observée de cette variable est "aberrante" par rapport à sa loi sous $\mathcal{H}_0$, on décidera de rejeter l'hypothèse.

6.1 Définitions générales

On observe un échantillon iid $X_1, ..., X_n$ de loi μ dont on veut tester une ou plusieurs propriétés (par exemple, tester si μ est une loi normale, μ est centrée etc...). De manière générale, on s'intéresse à tester une hypothèse de la forme $\mathcal{H}_0 : \mu \in \mathcal{M}_0$ pour $\mathcal{M}_0$ un certain sous-ensemble du modèle initial $\mathcal{M}$. L'hypothèse $\mathcal{H}_0$ est appelée hypothèse nulle. L'ensemble des possibilités si $\mathcal{H}_0$ n'est pas vérifiée est l'hypothèse alternative, notée $\mathcal{H}_1$. On a donc $\mathcal{H}_1 : \mu \in \mathcal{M}_1 = \mathcal{M} \setminus \mathcal{M}_0$.

Définition 6.1. *Un test de niveau α d'une hypothèse $\mathcal{H}_0$ est une statistique $\Phi_\alpha(X_1, ..., X_n)$ qui suit une loi de Bernoulli de paramètre α sous l'hypothèse $\mathcal{H}_0$.*

Il est indispensable que la variable aléatoire Φ_α soit une fonction connue des observations, c'est-à-dire que l'on peut observer une réalisation de Φ_α. C'est précisément la valeur observée de Φ_α qui va nous permettre de juger de la crédibilité de l'hypothèse $\mathcal{H}_0$.

Définition 6.2. *On appelle* p-value *du test le plus petit niveau α pour lequel on rejette $\mathcal{H}_0$ au vu de la réalisation de Φ_α:*

$$\text{p-value} = \inf\{\alpha > 0 : \Phi_\alpha(x_1, ..., x_n) = 1\}.$$

On rejettera donc l'hypothèse $\mathcal{H}_0$ si et seulement si la p-value est supérieure au niveau du test. L'écart entre la p-value et α quantifie la marge avec laquelle on accepte $\mathcal{H}_0$ dans le cas où elle est plus grande que α et la marge avec laquelle on rejette $\mathcal{H}_0$ dans le cas contraire.

Définition 6.3. *La puissance du test, notée $1 - \beta(.)$, est la probabilité de rejeter $\mathcal{H}_0$ sous une loi ν vérifiant l'hypothèse alternative*

$$1 - \beta(\nu) = \mathbb{E}_\nu\big(\Phi_\alpha(X_1, ..., X_n)\big) \ , \ \nu \in \mathcal{M}_1.$$

La puissance est une fonction de la loi ν, définie sur l'ensemble $\mathcal{M}_1$. La quantité $\beta(\nu)$ est la probabilité de ne pas rejeter $\mathcal{H}_0$ si l'échantillon $X_1, ..., X_n$ est issu de la loi ν. On dira qu'un test est convergent si la puissance tend vers 1 quand $n \to \infty$, pour tout $\nu \in \mathcal{M}_1$.

Il y a deux types d'erreurs possibles lors d'un test statistique:

- Rejeter $\mathcal{H}_0$ alors qu'elle est vraie constitue l'erreur de première espèce, de probabilité égale au niveau α du test.
- Ne pas rejeter $\mathcal{H}_0$ lorsque l'échantillon est issu d'une loi $\nu \in \mathcal{M}_1$ constitue l'erreur de deuxième espèce, de probabilité $\beta(\nu)$.

6.2 Tests paramétriques

La construction d'un test paramétrique est étroitement liée aux intervalles de confiance. Soit un paramètre θ, supposons que l'on cherche à tester une hypothèse de la forme $\mathcal{H}_0 : \theta = \theta_0$ où θ_0 est une valeur fixée par le statisticien. L'idée est de construire une variable aléatoire à partir des observations qui, si $\mathcal{H}_0$ est vraie, a une loi connue et de regarder si la valeur de la réalisation de la variable aléatoire est compatible avec l'hypothèse.

Exemples

1. Expérience de Bernoulli: on souhaite tester l'hypothèse $p = \frac{1}{2}$. Si $\mathcal{H}_0$ est vraie, alors on sait que la variable

$$T = 2\sqrt{n}\Big(\hat{p} - \frac{1}{2}\Big),$$

converge en loi vers une loi normale $\mathcal{N}(0,1)$. Autrement dit, si l'hypothèse $\mathcal{H}_0$ est vraie (et n est suffisamment grand), T appartient à l'intervalle $[-1,96, 1,96]$ avec probabilité $\approx 95\%$. Pour tester cette hypothèse, on calcule la valeur de T obtenue à partir des observations (on la note $t = T(x_1, ..., x_n)$). Si la valeur observée de T appartient à l'intervalle $[-1,96, 1,96]$ alors l'hypothèse est raisonnable. Si T prend une valeur hors de l'intervalle, alors l'hypothèse $\mathcal{H}_0$ n'est pas raisonnable et on peut choisir de rejeter cette hypothèse, avec néanmoins une probabilité de 5% de rejeter l'hypothèse alors qu'elle est vraie. La marge d'erreur est appelée le *niveau* du test, le choix $\alpha = 5\%$ est le plus courant.

La p-value du test est le quantile correspondant à la valeur observée t de la statistique de test, elle vérifie

$$1 - \text{p-value} = \mathbb{P}(-|t| \leq \mathcal{N}(0,1) \leq |t|).$$

2. Test sur une moyenne dans un modèle Gaussien: on veut tester l'hypothèse $m = m_0$ dans le modèle Gaussien où les observations $X_1, ..., X_n$ sont de loi $\mathcal{N}(\mathsf{m}, \sigma^2)$. Pour commencer, on supposera σ^2 connue. On sait que, si $\mathcal{H}_0$ est vraie

$$T = \sqrt{n}\ \frac{\hat{\mathsf{m}} - \mathsf{m}_0}{\sigma} \sim \mathcal{N}(0,1).$$

Ainsi, observer une valeur de T qui n'est pas dans l'intervalle $[-1,96; 1,96]$ est peu probable si $\mathcal{H}_0$ est vraie (la probabilité serait de 5% ce qui correspond à la marge d'erreur fixée à l'avance). On décide donc de rejeter $\mathcal{H}_0$ si T est hors de l'intervalle. Si la valeur observée de T est dans l'intervalle, alors on ne rejette pas l'hypothèse $\mathcal{H}_0$ au risque 5% (ce qui ne signifie pas pour autant qu'on accepte l'hypothèse $\mathsf{m} = \mathsf{m}_0$).
Dans le cas où la variance est inconnue, alors on construit la statistique

$$T = \sqrt{n}\ \frac{\hat{\mathsf{m}} - \mathsf{m}_0}{\hat{\sigma}} \quad \text{où} \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mathsf{m}})^2,$$

qui suit une loi de Student à $n-1$ degré de liberté sous $\mathcal{H}_0$. On peut appliquer le même raisonnement avec un intervalle approprié correspondant à la loi de Student $\mathcal{T}(n-1)$.

6.3 Tests non paramétriques

Les tests dits *non-paramétriques* s'appliquent à des modèles de dimension infinie et donc ne nécessitent pas de contraintes fortes sur la loi des observations. Du fait de leur robustesse, ces tests ne sont généralement valables qu'asymptotiquement.

6.3.1 Test sur la moyenne

On s'intéresse à tester une hypothèse de la forme $\mathcal{H}_0 : \mathbb{E}(X) = \mathsf{m}_0$ où m_0 est un réel connu. Le modèle considéré est l'ensemble des lois non constantes qui admettent un moment d'ordre deux

$$\mathcal{M} = \{\nu : 0 < \text{var}_\nu(X) < \infty\}.$$

L'ensemble des lois vérifiant l'hypothèse nulle est $\mathcal{M}_0 = \{\nu : \mathbb{E}_\nu(X) = \mathsf{m}_0,\ 0 < \text{var}_\nu(X) < \infty\}$ et l'hypothèse alternative est simplement $\mathcal{H}_1 : \mathbb{E}_\mu(X) \neq \mathsf{m}_0,\ 0 < \text{var}_\mu(X) < \infty$.

On définit la moyenne empirique $\overline{X}_n$ et l'écart-type empirique $\hat{\sigma}$ par

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Le test non-paramétrique sur la moyenne est basé sur le théorème central limite, et plus précisément, sur le comportement de la statistique

$$S_n := \sqrt{n}\ \frac{\overline{X}_n - \mathsf{m}_0}{\hat{\sigma}},$$

qui, si $\mathcal{H}_0$ est vraie, converge vers une loi normale standard quand n tend vers l'infini. Soit q_α le quantile d'ordre $\alpha \in]0,1[$ de la loi normale standard, le test est défini par

$$\Phi_\alpha(X_1, ..., X_n) = \mathbb{1}\left\{|S_n| > q_{1-\frac{\alpha}{2}}\right\}.$$

On choisit donc de rejeter $\mathcal{H}_0 : \mathbb{E}(X) = \mathsf{m}_0$ si la statistique $|S_n|$ prend une valeur trop grande pour une loi normale standard.

Théorème 6.4. *Le test $\Phi_\alpha = \mathbb{1}\{|S_n| > q_{1-\frac{\alpha}{2}}\}$ est asymptotiquement de niveau α,*

$$\forall \nu \in \mathcal{M}_0,\ \lim_{n\to\infty} \mathbb{E}_\nu(\Phi_\alpha) = \alpha.$$

De plus, $\forall \nu \in \mathcal{M}_1,\ \lim_{n\to\infty} \mathbb{E}_\nu(\Phi_\alpha) = 1$ (le test est convergent).

Preuve. On montre facilement par la loi des grands nombres que $\hat{\sigma}^2$ converge en probabilité (et même presque sûrement) vers $\sigma^2 = \mathrm{var}(X) > 0$. Par le théorème central limite, on sait que

$$Y_n := \sqrt{n}\ \frac{\overline{X}_n - \mathbb{E}(X)}{\sigma} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

D'après le théorème de Slutsky, on a donc

$$Y_n \times \frac{\sigma}{\hat{\sigma}} \xrightarrow[n\to\infty]{loi} \mathcal{N}(0,1).$$

Sous $\mathcal{H}_0$, $\mathbb{E}(X) = \mathsf{m}_0$ et $S_n = Y_n \times \frac{\sigma}{\hat{\sigma}}$ converge donc en loi vers une $\mathcal{N}(0,1)$. Comme la fonction de répartition de la loi normale est continue, la convergence en loi implique la convergence simple des fonctions de répartition et

$$\lim_{n\to\infty} \mathbb{P}\big(|S_n| > q_{1-\frac{\alpha}{2}}\big) = \mathbb{P}\big(|\mathcal{N}(0,1)| > q_{1-\frac{\alpha}{2}}\big) = \alpha.$$

Sous $\mathcal{H}_1$, posons $\mathsf{m}_1 = \mathbb{E}(X) \neq \mathsf{m}_0$. On a

$$S_n = \sqrt{n}\ \frac{\overline{X}_n - \mathsf{m}_0}{\hat{\sigma}} = \sqrt{n}\ \frac{\overline{X}_n - \mathsf{m}_1}{\hat{\sigma}} + \sqrt{n}\ \frac{\mathsf{m}_1 - \mathsf{m}_0}{\hat{\sigma}} \underset{n\to\infty}{\approx} \mathcal{N}(0,1) + \sqrt{n}\ \frac{\mathsf{m}_1 - \mathsf{m}_0}{\sigma}$$

et il est clair que $\mathbb{P}\big(|S_n| > q_{1-\frac{\alpha}{2}}\big) \to 1$. □

Le test de l'hypothèse $\mathcal{H}_0 : \mathbb{E}(X) = \mathsf{m}_0$ contre $\mathcal{H}_1 : \mathbb{E}(X) \neq \mathsf{m}_0$ est un test bilatéral. Si on sait à l'avance que l'espérance est forcément supérieure ou égale à m_0 (resp. inférieure ou égale), il est plus efficace d'utiliser un test unilatéral avec comme alternative $\mathcal{H}_1 : \mathbb{E}(X) > \mathsf{m}_0$ (resp. $\mathcal{H}_1 : \mathbb{E}(X) < \mathsf{m}_0$). Le test unilatéral est alors défini par $\Phi_\alpha = \mathbb{1}\{S_n > q_{1-\alpha}\}$ (resp. $\Phi_\alpha = \mathbb{1}\{S_n < q_\alpha\}$).

6.3.2 Test du signe

Le test sur la médiane (parfois appelé test du signe) est un des plus vieux tests non-paramétriques existant. Il permet de tester si la médiane de X, $\mathrm{med}(X) = F^-(0.5)$, est égale à une certaine valeur m_0. Comme la médiane n'est pas forcément unique, l'hypothèse nulle que l'on considère ici est

$$\mathcal{H}_0 : F(\mathsf{m}_0) = 0.5,$$

ce qui revient à tester, dans le cas où F est continue, si m_0 est un quantile d'ordre 0.5 de X (cf. la définition 2.14). On utilise pour ce test la statistique

$$M_n = \sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\} - \sum_{i=1}^{n} \mathbb{1}\{X_i < \mathsf{m}_0\} = \#\{i : X_i > \mathsf{m}_0\} - \#\{i : X_i < \mathsf{m}_0\}.$$

Comme pour le test sur la moyenne, on peut utiliser un test bilatéral ou unilatéral. Soit q_α le quantile d'ordre $\alpha \in]0,1[$ de la loi normale centrée réduite, on construit les tests

$$\begin{aligned}
&\Phi_\alpha = \mathbb{1}\{|M_n| > \sqrt{n} q_{1-\alpha/2}\} && \text{pour tester } \mathcal{H}_0 \text{ contre } \mathcal{H}_1 : F(\mathsf{m}_0) \neq 0.5 \\
&\Phi_\alpha = \mathbb{1}\{M_n > \sqrt{n} q_{1-\alpha}\} && \text{pour tester } \mathcal{H}_0 \text{ contre } \mathcal{H}_1 : F(\mathsf{m}_0) < 0.5 \\
&\Phi_\alpha = \mathbb{1}\{M_n < \sqrt{n} q_{\alpha}\} && \text{pour tester } \mathcal{H}_0 \text{ contre } \mathcal{H}_1 : F(\mathsf{m}_0) > 0.5
\end{aligned}$$

Théorème 6.5. *Si F est continue en m_0, ces trois tests sont de niveau α asymptotiquement et sont convergents.*

Preuve. Comme F est continue en m_0, $\mathbb{P}(X_i = \mathsf{m}_0) = 0$ et $M_n = 2\sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\} - n$ presque sûrement. Or, par le théorème central limite,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\} - \mathbb{P}(X > \mathsf{m}_0)\right) \xrightarrow[n\to\infty]{loi} \mathcal{N}\Big(0, \frac{1}{4}\Big).$$

Sous $\mathcal{H}_0$, $\mathbb{P}(X > \mathsf{m}_0) = 0.5$ et $M_n/\sqrt{n}$ converge vers une loi normale standard. Sous l'hypothèse alternative, il est facile de voir en distingant les cas, que $M_n/\sqrt{n}$ diverge et que le test est convergent. □

Remarque 6.6. *La statistique $\sum_{i=1}^{n} \mathbb{1}\{X_i > \mathsf{m}_0\}$ suit une loi binomiale $\mathcal{B}(n, 0.5)$ sous $\mathcal{H}_0$, on peut donc utiliser les quantiles de la loi binomiale pour construire un test plus précis pour des petites valeurs de n.*

6.4 Tests d'adéquation

Un test d'adéquation est un test sur une valeur particulière de la loi des observations. Etant donné une loi μ_0 connue, on veut tester si l'échantillon $X_1, ..., X_n$ est issu de cette loi.

6.4.1 Test du χ^2

Le test d'adéquation du χ^2 est utilisé à l'origine pour tester si un échantillon est issu d'une loi discrète μ_0. L'hypothèse nulle s'écrit donc $\mathcal{H}_0 : \mu = \mu_0$, pour μ_0 est une loi discrète connue à support fini $\{t_1, ..., t_k\}$. On se place dans le modèle non-paramétrique, l'hypothèse alternative est $\mathcal{H}_1 : \mu \neq \mu_0$. En notant $\pi_j = \mu_0(\{t_j\})$ et $p_j = \mathbb{P}(X = t_j)$ pour $j = 1, ..., k$, l'hypothèse nulle peut s'écrire de façon équivalente $\mathcal{H}_0 : p_j = \pi_j,\ j = 1, ..., k$. On remarque également que l'hypothèse alternative peut s'écrire $\mathcal{H}_1 : \exists j,\ p_j \neq \pi_j$.

On estime naturellement les probabilités inconnues p_j par les fréquences d'apparition des valeurs t_j dans l'échantillon

$$p_{nj} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i = t_j\}\ ,\ j = 1, ..., k.$$

On définit alors la statistique de test

$$Q_n := n\sum_{j=1}^{k} \frac{(p_{nj} - \pi_j)^2}{\pi_j} = n\Big(\sum_{j=1}^{k} \frac{p_{nj}^2}{\pi_j} - 1\Big).$$

Proposition 6.7. *Sous $\mathcal{H}_0 : \mu = \mu_0$, Q_n converge en loi vers un $\chi^2(k-1)$ quand $n \to \infty$. Par ailleurs, Q_n tend vers l'infini en probabilité sous $\mathcal{H}_1 : \mu \neq \mu_0$.*

Preuve. On pose $p_n = (p_{n1}, ..., p_{nk})^\top$, $\pi = (\pi_1, ..., \pi_k)^\top$ et

$$Y_i = \begin{bmatrix} \mathbb{1}\{X_i = t_1\} \\ \vdots \\ \mathbb{1}\{X_i = t_k\} \end{bmatrix}, \ i = 1, ..., n.$$

Par le théorème central limite vectoriel, $\sqrt{n}\big(p_n - \pi\big) = \sqrt{n}\ \frac{1}{n}\sum_{i=1}^n (Y_i - \pi) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Sigma)$ où

$$\Sigma = \mathrm{var}(Y) = \begin{bmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_k \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & \dots & -\pi_2\pi_k \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_1\pi_k & -\pi_2\pi_k & \dots & \pi_k(1-\pi_k) \end{bmatrix}.$$

Soit $\mathsf{D} = \mathrm{Diag}(\sqrt{\pi})$, la matrice diagonale de diagonale $\sqrt{\pi} = (\sqrt{\pi_1}, ..., \sqrt{\pi_k})^\top$. On déduit du résultat précédent que

$$q_n := \sqrt{n}\ \mathsf{D}^{-1}\big(p_n - \pi\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \Gamma),$$

où $\Gamma = \mathsf{D}^{-1}\Sigma\mathsf{D}^{-1} = \mathrm{I} - \sqrt{\pi}\sqrt{\pi}^\top$ (I désigne l'identité de $\mathbb{R}^k$). On remarque dans un premier temps que q_n est orthogonal à $\sqrt{\pi}$. Soit $\{\mathsf{v}_1, ..., \mathsf{v}_{k-1}\}$ une collection de vecteurs orthonormés de $\mathbb{R}^k$, orthogonaux à $\sqrt{\pi}$ et $\mathsf{V} = (\mathsf{v}_1, ..., \mathsf{v}_{k-1}) \in \mathbb{R}^{k\times(k-1)}$. Comme la norme Euclidienne est invariante par changement de base othonormée, on déduit

$$Q_n = \|q_n\|^2 = (q_n^\top\sqrt{\pi})^2 + \|\mathsf{V}^\top q_n\|^2 = \|\mathsf{V}^\top q_n\|^2.$$

En remarquant que $\mathsf{V}^\top\Gamma\mathsf{V} = \mathrm{I}$ (l'identité de $\mathbb{R}^{k-1}$), on déduit

$$\mathsf{V}^\top q_n = \sqrt{n}\ \mathsf{V}^\top\mathsf{D}^{-1}\big(p_n - \pi\big) \xrightarrow[n\to\infty]{loi} \mathcal{N}(0, \mathrm{I}),$$

d'où le résultat $Q_n \xrightarrow[n\to\infty]{loi} \chi^2(k-1)$. □

Ce résultat permet de construire le test d'adéquation du χ^2 qui s'écrit

$$\Phi_\alpha = \mathbb{1}\{Q_n > q_{1-\alpha}\},$$

où $q_{1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de la loi du $\chi^2(k-1)$. Par la continuité de la fonction de répartition d'un χ^2, on déduit facilement de la proposition précédente que le test est asymptotiquement de niveau α et qu'il est convergent.

L'idée du test d'adéquation du χ^2 peut se généraliser à tout type de loi. Supposons maintenant que les observations sont distribuées suivant une loi μ quelconque sur $\mathbb{R}$ et que l'on veut tester une valeur particulière $\mathcal{H}_0 : \mu = \mu_0$. On procède de la façon suivante:

- On choisit une partition $\{A_j\}_{j=1,...,k}$ de $\mathbb{R}$ telle que $\pi_j := \mu_0(A_j) > 0$ pour tout $j = 1, .., k$.
- On calcule les proportions d'observations pour chaque ensemble A_j,

$$p_{nj} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \in A_j\}, \ j = 1, ..., k.$$

- D'après la proposition 6.7, on sait que

$$Q_n = n \sum_{j=1}^{k} \frac{(p_{nj} - \pi_j)^2}{\pi_j}$$

suit asymptotiquement un $\chi^2(k-1)$ si $\mu(A_j) = \pi_j$ pour tout $j = 1, .., k$.

Remarque 6.8. *Le test $\Phi_\alpha = \mathbb{1}\{Q_n > q_{1-\alpha}\}$ est donc asymptotiquement de niveau α pour tester l'hypothèse $\mathcal{H}_0 : \mu = \mu_0$. En revanche, il n'est plus convergent puisqu'on peut construire une loi $\mu_1 \neq \mu_0$ telle que $\mu_1(A_j) = \pi_j$ pour tout $j = 1, .., k$. La puissance du test sous μ_1 convergera donc vers $1 - \alpha$, ce qui n'est pas satisfaisant. Une solution à ce problème est d'augmenter le nombre de classes $k = k_n$ de la partition avec n, suffisamment lentement pour pouvoir approcher la loi de la statistique Q_n par un $\chi^2(k_n)$.*

6.4.2 Test de Kolmogorov-Smirnov

Comme le test d'adéquation du χ^2, le test de Kolmogorov-Smirnov permet de tester une valaur particulière de la loi via l'hypothèse nulle $\mathcal{H}_0 : \mu = \mu_0$. En revanche, il n'est valable que pour des lois μ_0 continues. L'idée du test repose sur le théorème suivant.

Théorème 6.9 (Donsker)**.** *Soit $X_1, ..., X_n$ un échantillon iid de loi continue de fonction de répartition F et F_n la fonction de répartition empirique. Alors,*

$$K_n = \sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n \to \infty]{loi} K,$$

où K est une variable aléatoire de loi de Kolmogorov de fonction de répartition

$$F_K(x) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 x^2), \ , x \geq 0.$$

La preuve de ce résultat est une conséquence d'un résultat plus général concernant la convergence du processus empirique renormalisé. Précisément, dans le cas de variables aléatoires X_i uniformes sur $[0, 1]$, le processus $\sqrt{n}\big(F_n(t) - F(t)\big), t \in [0, 1]$ converge en loi vers un pont Brownien dans l'espace des fonctions càdlàg muni de la topologie de Skorokhod.

La fonction de répartition de la loi de Kolmogorov est continue est strictement croissante sur $]0, +\infty[$, il existe donc un unique $q_{1-\alpha}$ tel que $F_K(q_{1-\alpha}) = 1 - \alpha$, ce qui permet de construire un test asymptotiquement de niveau α par $\Phi_\alpha = \mathbb{1}\{K_n > q_{1-\alpha}\}$. On peut montrer également que ce test est convergent pour toute fonction de répartition $F_1 \neq F$.

7 Modèle linéaire

On cherche maintenant à modéliser le lien entre une variable Y aléatoire et plusieurs variables explicatives $X_1, ..., X_p$ considérées ici comme déterministes. On suppose que les observations $\{y_i, x_{1i}, ..., x_{pi}\}_{i=1,...,n}$ vérifient

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} + \epsilon_i, \ i = 1, ..., n,$$

où les ϵ_i sont centrés, de même variance $\sigma^2 < \infty$ et non-corrélés. Le modèle s'écrit sous forme matricielle

$$y = \mathsf{X}\beta + \epsilon,$$

avec

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ \mathsf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \ \text{ et } \ \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

La matrice X est supposée injective ce qui implique en particulier que $n > p$. Comme on suppose les variables explicatives x_i déterministes (non aléatoires), l'alea est uniquement dû à la présence des bruits ϵ_i. Ainsi, seuls les vecteur y et ϵ sont des réalisations des variables aléatoires, X et β sont déterministes (par ailleurs, seuls y et X sont connus du statisticien). On suppose de plus que les bruits ϵ_i sont

- centrés: $\mathbb{E}(\epsilon_i) = 0$,
- non-corrélés: $\forall i \neq j, \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$,
- de variances égales (homoscédastiques): $\mathrm{var}(\epsilon_i) = \sigma^2 < \infty$.

Ces hypothèses, dites *hypothèses faibles* du modèle linéaire, sont résumées par les égalités matricielles $\mathbb{E}(\epsilon) = 0$ et $\mathrm{var}(\epsilon) = \sigma^2 \mathrm{I}$. Les *hypothèses fortes* du modèle linéaires supposent en plus que ϵ est un vecteur Gaussien, auquel cas les ϵ_i sont indépendants car non-corrélés.

7.1 Méthode des moindres carrés ordinaires

On cherche à estimer β et σ^2. L'estimateur des MCO $\hat{\beta}$ est défini comme l'unique minimiseur de

$$R(\mathsf{b}) = \|y - \mathsf{X}\mathsf{b}\|^2, \quad \mathsf{b} \in \mathbb{R}^{p+1}.$$

On rappelle que $p < n$ et que X est de rang $p+1$. Sous ces hypothèses, l'estimateur $\hat{\beta}$ est l'unique solution des conditions du premier ordre

$$\nabla R(\hat{\beta}) = -2\mathsf{X}^\top y + 2\mathsf{X}^\top \mathsf{X}\hat{\beta} = 0 \iff \hat{\beta} = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top y.$$

On remarque que $\mathsf{X}^\top \mathsf{X}$ est inversible du fait que X est injective.

Proposition 7.1. *L'estimateur des MCO $\hat{\beta}$ est un estimateur sans biais de β de matrice de variance*

$$\mathrm{var}(\hat{\beta}) = \sigma^2 (\mathsf{X}^\top \mathsf{X})^{-1}.$$

Preuve. Comme X est déterministe, on a $\mathbb{E}(y) = \mathbb{E}(\mathsf{X}\beta + \epsilon) = \mathsf{X}\beta$ et $\mathrm{var}(y) = \mathrm{var}(\epsilon) = \sigma^2 \mathrm{I}$. On obtient

$$\begin{aligned} \mathbb{E}(\hat{\beta}) &= \mathbb{E}((\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top y) = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top \mathbb{E}(y) = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top \mathsf{X}\beta = \beta \\ \mathrm{var}(\hat{\beta}) &= \mathrm{var}((\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top y) = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top \, \mathrm{var}(y) \, \mathsf{X}(\mathsf{X}^\top \mathsf{X})^{-1} = \sigma^2 (\mathsf{X}^\top \mathsf{X})^{-1}. \end{aligned}$$

□

Théorème 7.2. *(Gauss-Markov) L'estimateur des moindres carrés $\hat{\beta}$ est optimal (au sens du coût quadratique) parmi les estimateurs sans biais linéaires en y.*

Preuve. Soit $\tilde{\beta}$ un estimateur sans biais de β, linéaire en y. On a donc $\tilde{\beta} = \mathsf{A}y$ pour une matrice déterministe $\mathsf{A} \in \mathbb{R}^{(p+1)\times n}$ telle que,

$$\forall \beta \in \mathbb{R}^{p+1} \ , \ \beta = \mathbb{E}(\mathsf{A}y) = A\mathbb{E}(y) = \mathsf{A}\mathsf{X}\beta \iff \mathsf{A}\mathsf{X} = \mathrm{I}\,.$$

On veut montrer que la matrice $\mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta})$ est semi définie-positive. On a

$$\begin{aligned}\mathrm{var}(\tilde{\beta}) &= \mathsf{A}\,\mathrm{var}(y)\mathsf{A}^\top = \sigma^2\mathsf{A}\mathsf{A}^\top = \sigma^2\left[\mathsf{A}\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top\mathsf{A}^\top + \mathsf{A}(\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{A}^\top\right]\\ &= \sigma^2(\mathsf{X}^\top\mathsf{X})^{-1} + \sigma^2\mathsf{A}(\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{A}^\top = \mathrm{var}(\hat{\beta}) + \sigma^2\mathsf{A}(\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{A}^\top.\end{aligned}$$

Il reste à remarquer que la matrice $\mathsf{A}(\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{A}^\top = \mathrm{var}(\tilde{\beta}) - \mathrm{var}(\hat{\beta})$ est semi définie-positive. En effet, $\forall \mathsf{a} \in \mathbb{R}^{p+1}$,

$$\mathsf{a}^\top\mathsf{A}(\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{A}^\top\mathsf{a} = \|(\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\mathsf{A}^\top\mathsf{a}\|^2 \geq 0. \qquad \square$$

L'optimalité au sens $\mathbb{L}^2$ parmi les estimateurs linéaires sans biais ne nécessite pas la normalité du modèle. Un résultat plus fort est valable dans le cas Gaussien $\epsilon \sim \mathcal{N}(0, \sigma^2\,\mathrm{I})$ où la variance de $\hat{\beta}$ atteint la borne de Cramer-Rao. L'estimateur des moindres carrés est donc optimal parmi tous les estimateurs sans biais de β dans ce cas.

La matrice $\Pi_{\mathsf{X}} := \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top$ utilisée dans la preuve du théorême de Gauss-Markov est la projection orthogonale sur l'image de X. On le montre simplement en vérifiant que Π_{X} est symétrique et vérifie $\Pi_{\mathsf{X}}^2 = \Pi_{\mathsf{X}}$ et $\mathrm{Im}(\Pi_{\mathsf{X}}) = \mathrm{Im}(\mathsf{X})$. Ainsi, le vecteurs des prévisions

$$\hat{y} = \mathsf{X}\hat{\beta} = \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top y,$$

est la projection orthogonale de y sur $\mathrm{Im}(\mathsf{X})$. C'est en quelque sorte la part de $y \in \mathbb{R}^n$ expliquée par les variables $\mathbf{1}, x_1, ..., x_p$ (les colonnes de X). De même, le vecteur des résidus

$$\hat{\epsilon} = y - \hat{y} = (\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)y = (\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)(\mathsf{X}\beta + \epsilon) = (\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top)\epsilon$$

est la projection orthogonale de y sur $\mathrm{Im}(\mathsf{X})^\perp$ et par conséquent celle de ϵ puisque $y = \mathsf{X}\beta + \epsilon$. Une conséquence immédiate est que les vecteurs $\hat{\beta}$ et $\hat{\epsilon}$ sont non-corrélés. En effet,

$$\mathrm{cov}(\hat{\beta}, \hat{\epsilon}) = \mathbb{E}\left[(\hat{\beta} - \beta)\hat{\epsilon}^\top\right] = (\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top\mathbb{E}\left[\epsilon\epsilon^\top\right](\mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top) = 0.$$

La norme du vecteur des résidus permet de construire un estimateur de σ^2 par

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|y - \hat{y}\|^2 = \frac{1}{n-p-1}\|\hat{\epsilon}\|^2.$$

Proposition 7.3. *L'estimateur $\hat{\sigma}^2$ est sans biais.*

Preuve. On a vu que $\hat{\epsilon} = \Pi_{\mathsf{X}^\perp}\epsilon$ où $\Pi_{\mathsf{X}^\perp} = \mathrm{I}-\mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top$ est la matrice de projection orthogonale sur $\mathrm{Im}(\mathsf{X})^\perp$. On utilise qu'un réél est égal à sa trace et que $\mathrm{tr}(AB) = \mathrm{tr}(BA)$:

$$\mathbb{E}\|\hat{\epsilon}\|^2 = \mathbb{E}(\epsilon^\top\Pi_{\mathsf{X}^\perp}^\top\Pi_{\mathsf{X}^\perp}\epsilon) = \mathbb{E}(\epsilon^\top\Pi_{\mathsf{X}^\perp}\epsilon) = \mathbb{E}\left[\mathrm{tr}(\epsilon^\top\Pi_{\mathsf{X}^\perp}\epsilon)\right] = \mathbb{E}\left[\mathrm{tr}(\Pi_{\mathsf{X}^\perp}\epsilon\epsilon^\top)\right].$$

Clairement, la trace (somme des éléments diagonaux) commute avec l'espérance, d'où

$$\mathbb{E}\|\hat{\epsilon}\|^2 = \mathrm{tr}\left[\Pi_{\mathsf{X}^\perp}\mathbb{E}(\epsilon\epsilon^\top)\right] = \mathrm{tr}\left[\Pi_{\mathsf{X}^\perp}\sigma^2\,\mathrm{I}\right] = \sigma^2\,\mathrm{tr}(\Pi_{\mathsf{X}^\perp}).$$

La trace (somme des valeurs propres) d'une matrice de projection étant égale à son rang, on obtient

$$\mathbb{E}\|\hat{\epsilon}\|^2 = \sigma^2(n-p-1) \iff \mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2. \qquad \square$$

Un résultat plus fort est valable dans le cas Gaussien.

Proposition 7.4. *Dans le modèle Gaussien $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$, les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants et vérifient*

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathsf{X}^\top \mathsf{X})^{-1}) \quad \textit{et} \quad (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1).$$

Preuve. On a $y = \mathsf{X}\beta + \epsilon \sim \mathcal{N}(\mathsf{X}\beta, \sigma^2 \mathrm{I})$, d'où $\hat{\beta} = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top y \sim \mathcal{N}(\beta, \sigma^2(\mathsf{X}^\top \mathsf{X})^{-1})$. De même, $\hat{\epsilon} = \Pi_{\mathsf{X}^\perp}\epsilon \sim \mathcal{N}(0, \sigma^2 \Pi_{\mathsf{X}^\perp})$. En exprimant $\hat{\epsilon}$ dans une base orthonormée de $\mathrm{Im}(\mathsf{X})^\perp$, on déduit (c'est une conséquence du théorême de Cochran)

$$\frac{\|\hat{\epsilon}\|^2}{\sigma^2} = (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1).$$

Il reste à montrer l'indépendance entre $\hat{\beta}$ et $\hat{\sigma}^2$. On a vu que $\hat{\beta}$ et $\hat{\epsilon}$ sont non-corrélés et donc indépendants dans le cas Gaussien. Ainsi, $\hat{\sigma}^2$ qui est fonction de $\hat{\epsilon}$ est indépendant de $\hat{\beta}$. □

Remarque 7.5. *Dans le cas Gaussien, $\hat{\beta}$ est l'estimateur du maximum de vraisemblance. En revanche, l'estimateur du maximum de vraismeblance de σ^2 est différent, donné par*

$$\hat{\sigma}^2_{\mathrm{MV}} = \frac{1}{n}\|y - \hat{y}\|^2 = \frac{n-p-1}{n}\hat{\sigma}^2.$$

7.2 Comportement asymptotique des estimateurs

On s'intéresse maintenant au comportement des estimateurs quand n tend vers l'infini. Sous les hypothèses fortes de la régression, les lois de $\hat{\beta}$ et $\hat{\sigma}^2$ sont connues ce qui permet de déduire facilement leur comportement asymptotique. La convergence de $\hat{\beta}$ dans $\mathbb{L}^2$ est soumise à la seule condition que $(\mathsf{X}^\top \mathsf{X})^{-1}$ tend vers 0. La convergence de $\hat{\sigma}^2$, que ce soit dans $\mathbb{L}^2$ ou même presque sûrement, est vérifiée dans le cas Gaussien sans hypothèse supplémentaire sur X (c'est une conséquence immédiate du théorême 3.4). On peut se demander si ces résultats restent valables sans la normalité des bruits. Un premier résultat immédiat est que sous les hypothèses faibles, $\hat{\beta}$ reste convergent dans $\mathbb{L}^2$ dès que $(\mathsf{X}^\top \mathsf{X})^{-1}$ tend vers 0. On peut également montrer que si les ϵ_i sont iid et $h_n := \max_{1 \le i,j \le p+1} |\Pi_{\mathsf{X},ij}|$ tend vers 0, alors $\hat{\beta}$ est asymptotiquement Gaussien. Si de plus $\frac{1}{n}\mathsf{X}^\top \mathsf{X}$ converge quand $n \to \infty$ vers une matrice inversible M, alors

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{loi} \mathcal{N}(0, \sigma^2 \mathsf{M}^{-1}).$$

L'hypothèse $\frac{1}{n}\mathsf{X}^\top \mathsf{X} \to \mathsf{M}$ est souvent vérifiée en pratique. Par exemple, elle est vérifiée presque sûrement si l'échantillon $\{x_{1i}, ..., x_{pi}\}_{i=1,...,n}$ est issu de réalisations indépendantes de variables aléatoires $X_1, ..., X_p$ de carré intégrable. Dans ce cas, $\frac{1}{n}\mathsf{X}^\top \mathsf{X}$ converge presque sûrement vers la matrice des moments d'ordre deux, par la loi forte des grands nombres. Ce résultat est important car il confirme que même sous les hypothèses faibles, la plupart des tests du modèle linéaire restent valables asymptotiquement.

On aura besoin du lemme suivant que l'on admettra.

Lemme 7.6. *Soit $\{\xi_i\}_{i \in \mathbb{N}}$ une suite de vecteurs aléatoires centrés, indépendants et tels que la matrice $\Sigma_n = \sum_{i=1}^n \mathrm{var}(\xi_i)$ est inversible pour n suffisamment grand.*

1. *Si $\Sigma_n = O(n)$, alors $\displaystyle\frac{1}{n}\sum_{i=1}^n \xi_i \xrightarrow{ps} 0$.*

2. *Si $\sup_{i \in \mathbb{N}} \mathbb{E}\|\xi_i\|^{2+\delta} < \infty$ pour un $\delta > 0$, alors $\displaystyle\Sigma_n^{-\frac{1}{2}} \sum_{i=1}^n \xi_i \xrightarrow{loi} \mathcal{N}(0, \mathrm{I})$.*

Théorème 7.7. *Si les ϵ_i sont indépendants, alors $\hat{\beta}$ converge presque sûrement. S'ils vérifient de plus $\sup_{i\in\mathbb{N}} \mathbb{E}|\epsilon_i|^{2+\delta} < \infty$ pour un $\delta > 0$, alors*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{loi} \mathcal{N}(0, \sigma^2 \mathsf{M}^{-1}).$$

Preuve. Soit X_i la i-ème ligne de la matrice X, on pose $\xi_i = \mathsf{X}_i^\top \epsilon_i$. D'après la propriété 1. du lemme,

$$\hat{\beta} = \beta + n(\mathsf{X}^\top\mathsf{X})^{-1}\ \frac{1}{n}\mathsf{X}^\top\epsilon = \beta + n(\mathsf{X}^\top\mathsf{X})^{-1}\ \frac{1}{n}\sum_{i=1}^n \xi_i \xrightarrow{ps} \beta,$$

où on a utilisé que $n(\mathsf{X}^\top\mathsf{X})^{-1}$ converge vers M^{-1}. Pour la convergence en loi, on utilise les conditions de Lyapunov (propriété 2. du lemme) appliquée cette fois aux vecteurs $\xi_i = (\mathsf{X}^\top\mathsf{X})^{-\frac{1}{2}}\mathsf{X}_i^\top\epsilon_i$. On a $\mathbb{E}\|\xi_i\|^{2+\delta} = [\mathsf{X}_i(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}_i^\top]^{1+\delta/2}\ \mathbb{E}|\epsilon_i|^{2+\delta} \leq \mathbb{E}|\epsilon_i|^{2+\delta}$ et

$$\Sigma_n = \sum_{i=1}^n \mathrm{var}(\xi_i) = \sigma^2(\mathsf{X}^\top\mathsf{X})^{-\frac{1}{2}}\Big(\sum_{i=1}^n \mathsf{X}_i^\top\mathsf{X}_i\Big)(\mathsf{X}^\top\mathsf{X})^{-\frac{1}{2}} = \sigma^2\,\mathrm{I}\,.$$

On déduit $\sum_{i=1}^n \xi_i \xrightarrow{loi} \mathcal{N}(0, \sigma^2\,\mathrm{I})$. Comme $\hat{\beta} - \beta = (\mathsf{X}^\top\mathsf{X})^{-\frac{1}{2}}\sum_{i=1}^n \xi_i$, on obtient

$$\sqrt{n}(\hat{\beta} - \beta) = \mathsf{M}^{-\frac{1}{2}}\sum_{i=1}^n \xi_i + \Big(\sqrt{n}(\mathsf{X}^\top\mathsf{X})^{-\frac{1}{2}} - \mathsf{M}^{-\frac{1}{2}}\Big)\sum_{i=1}^n \xi_i \xrightarrow{loi} \mathcal{N}(0, \sigma^2\mathsf{M}^{-1})$$

en utilisant que $\sqrt{n}(\mathsf{X}^\top\mathsf{X})^{-\frac{1}{2}} - \mathsf{M}^{-\frac{1}{2}} \to 0$. □

7.3 Analyse de la variance

L'analyse de la variance consiste à diviser la variance de y en une partie expliquée par les variables $x_1, ..., x_p$ et une partie résiduelle. Il s'agit de remarquer que

- $\hat{y} = \mathsf{X}(\mathsf{X}^\top\mathsf{X})^{-1}\mathsf{X}^\top y = \Pi_{\mathsf{X}} y$ est la projection orthogonale de y sur $\mathrm{Im}(\mathsf{X})$.
- $\overline{y}\mathbf{1}$ est la projection orthogonale de y sur l'espace engendré par le vecteur constant $\mathbf{1}$, noté $\mathrm{vec}\{\mathbf{1}\}$.
- $\mathrm{vec}\{\mathbf{1}\}$ étant un sous espace de $\mathrm{Im}(\mathsf{X})$, $\overline{y}\mathbf{1}$ est également la projection orthogonale de $\hat{y}$ sur $\mathrm{vec}\{\mathbf{1}\}$ (on le vérifie en remarquant que $\overline{\hat{y}} = \overline{y}$).
- $\hat{\epsilon} = y - \hat{y} = (\mathrm{I} - \Pi_{\mathsf{X}})y$ est la projection orthgonale de y sur $\mathrm{Im}(\mathsf{X})^\perp$. C'est également la projection orthgonale de ϵ sur $\mathrm{Im}(\mathsf{X})^\perp$ puisque $(\mathrm{I} - \Pi_{\mathsf{X}})\mathsf{X}\beta = 0$.

Ainsi, en décomposant $y - \overline{y}\mathbf{1} = \hat{y} - \overline{y}\mathbf{1} + \hat{\epsilon}$, le théorème de Pythagore nous donne

$$\underbrace{\|y - \overline{y}\mathbf{1}\|^2}_{\mathrm{SCT}} = \underbrace{\|\hat{y} - \overline{y}\mathbf{1}\|^2}_{\mathrm{SCE}} + \underbrace{\|\hat{\epsilon}\|^2}_{\mathrm{SCR}}.$$

Le coefficient de détermination R^2 qui donne un indicateur de la qualité de la modélisation est défini par

$$R^2 := \frac{\mathrm{SCE}}{\mathrm{SCT}} = \frac{\|\hat{y} - \overline{y}\mathbf{1}\|^2}{\|y - \overline{y}\mathbf{1}\|^2}.$$

Remarque 7.8. *Dans le cas univarié, on a vu que le coefficient de détermination est égal au carré du coefficient de corrélation de Pearson $\rho(x, y)$. Dans le cas multivarié, le R^2 correspond à la valeur maximale du carré du coefficient de Pearson entre y et une combinaison linéaire des variables explicatives:*

$$R^2 = \sup_{\mathsf{b}\in\mathbb{R}^{p+1}} \rho(y, \mathsf{X}\mathsf{b})^2 = \rho(y, \hat{y})^2.$$

7.4 Tests

Le coefficient de détermination permet de tester l'existence d'une relation linéaire entre y et les variables explicatives. Précisément, l'hypothèse nulle est l'absence de relation linéaire, ce qui se traduit par $\mathcal{H}_0$: $\beta_1 = ... = \beta_p = 0$ contre $\mathcal{H}_1 : \exists j, \beta_j \neq 0$. Ce test est parfois appelé *test de Fisher global.*

Proposition 7.9. *Dans le modèle Gaussien où $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$, la statistique*

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2} = \frac{n-p-1}{p} \frac{\mathrm{SCE}}{\mathrm{SCR}}$$

suit, sous $\mathcal{H}_0 : \beta_1 = ... = \beta_p = 0$, une loi de Fisher à p et $n-p-1$ degrés de liberté.

Preuve. On rappelle qu'une loi de Fisher à k_1 et k_2 degrés de liberté, que l'on note $\mathcal{F}(k_1, k_2)$, est le rapport de deux lois du χ^2 indépendantes renormalisées par leur nombres de degrés de liberté:

$$\mathcal{F}(k_1, k_2) \overset{loi}{=} \frac{\chi^2(k_1)/k_1}{\chi^2(k_2)/k_2}.$$

D'après la définition de R^2 et le fait que $SCT = \mathrm{SCE} + \mathrm{SCR}$, on a $1 - R^2 = \mathrm{SCR} / \mathrm{SCT}$ et

$$\frac{R^2}{1-R^2} = \frac{\mathrm{SCE}}{\mathrm{SCT}} \times \frac{\mathrm{SCT}}{\mathrm{SCR}} = \frac{\mathrm{SCE}}{\mathrm{SCR}} = \frac{\|\hat{y} - \overline{y}\mathbf{1}\|^2}{\|\hat{\epsilon}\|^2}.$$

Le vecteur $\hat{y} - \overline{y}\mathbf{1}$ est la projection orthogonale de y sur $E := \mathrm{Im}(\mathsf{X}) \cap \mathrm{vec}\{\mathbf{1}\}^\perp$. Sous $\mathcal{H}_0 : \beta_1 = ... = \beta_p = 0$, on a $y = \beta_0 \mathbf{1} + \epsilon$ et donc $\Pi_E y = \Pi_E \epsilon$. De plus, $\hat{\epsilon} = \Pi_{\mathsf{X}^\perp} \epsilon$. Les espaces E (de dimension p car X est de plein rang) et $\mathrm{Im}(\mathsf{X})^\perp$ (de dimension $n-p-1$) sont orthogonaux, les vecteurs $\hat{\epsilon}$ et $\hat{y} - \overline{y}\mathbf{1}$ sont donc indépendants et

$$\frac{\|\hat{\epsilon}\|^2}{\sigma^2} = \frac{\|\Pi_{\mathsf{X}^\perp} \epsilon\|^2}{\sigma^2} \sim \chi^2(n-p-1) \quad \text{et} \quad \frac{\|\hat{y} - \overline{y}\mathbf{1}\|^2}{\sigma^2} = \frac{\|\Pi_E \epsilon\|^2}{\sigma^2} \sim \chi^2(p). \qquad \square$$

Remarque 7.10. *Sous $\mathcal{H}_1 : \exists \beta_j \neq 0, 1 \leq j \leq p$, la statistique F tend vers l'infini en probabilité dès que $\|\Pi_E(\mathsf{X}\beta)\|$ tend vers l'infini (cette hypothèse est très raisonnable en pratique). Dans ce cas, le test est donc puissant asymptotiquement.*

La statistique F permet donc de tester s'il existe au moins une variable pertinente parmi les variables explicatives. Elle est calculée automatiquement dans la commande `summary(lm(`$y \sim x$`))` de `R`. Ce test reste valable asymptotiquement dans le cas non-Gaussien, sous des hypothèses raisonnables.
Evidemment, l'existence d'au moins une variable explicative pertinente n'implique pas forcément que toutes soient pertinentes. Pour tester individuellement chaque variable x_j, on peut utiliser un test de Student. Précisemment, on sait que dans le modèle Gaussien, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathsf{X}^\top \mathsf{X})^{-1})$, et en particulier,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathsf{X}^\top \mathsf{X})^{-1}]_{jj}}} \sim \mathcal{T}(n-p-1).$$

Ce résultat permet de tester l'hypothèse nulle $\mathcal{H}_0 : \beta_j = 0$ pour vérifier la pertinence de β_j (de manière générale, on peut tester une hypothèse de la forme $\mathcal{H}_0 : \beta_j = b$) et de construire des intervalles de confiance. La proposition suivante donne une procédure pour tester une hypothèse affine générale sur les paramètres.

Proposition 7.11. *Soient $\mathsf{A} \in \mathbb{R}^{q \times (p+1)}$ de rang $q \leq p+1$ et $\mathsf{b} \in \mathbb{R}^q$ connus. Sous les hypothèses fortes, la statistique*

$$F = \frac{(\mathsf{A}\hat{\beta} - \mathsf{b})^\top [\mathsf{A}(\mathsf{X}^\top \mathsf{X})^{-1} \mathsf{A}^\top]^{-1} (\mathsf{A}\hat{\beta} - \mathsf{b})}{q\hat{\sigma}^2}$$

suit, sous $\mathcal{H}_0 : \mathsf{A}\beta = \mathsf{b}$, une loi de Fisher $\mathcal{F}(q, n-p-1)$.

Preuve. On sait que par hypothèses $\mathsf{A}\hat{\beta} - \mathsf{b} \sim \mathcal{N}\left(\mathsf{A}\beta - \mathsf{b}, \sigma^2 \mathsf{A}(\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{A}^\top\right)$. On déduit que sous $\mathcal{H}_0$,

$$V = \frac{1}{\sigma}\,[\mathsf{A}(\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{A}^\top]^{-\frac{1}{2}}(\mathsf{A}\hat{\beta} - \mathsf{b})$$

est un vecteur Gaussien standard dans $\mathbb{R}^q$. Puisque $\hat{\beta}$ est indépendant de $\hat{\sigma}^2$, on conclut

$$F = \frac{\|V\|^2}{q} \times \frac{\sigma^2}{\hat{\sigma}^2} \sim \mathcal{F}(q, n-p-1). \qquad \square$$

Remarque 7.12. *Le test des hypothèses $\beta_1 = ... = \beta_p = 0$ ou encore $\beta_p = 0$ sont des cas particuliers du théorême, correspondant respectivement à $\mathsf{A} = (0, \mathrm{I}) \in \mathbb{R}^{p\times(p+1)}, \mathsf{b} = 0$ et $\mathsf{A} = (0, ..., 0, 1), \mathsf{b} = 0$. Les hypothèses de la forme $\beta_j = \beta_{j'}$ correspondent également à des valeurs particulières de A et b.*

Ce genre d'hypothèse sur les paramètres revient à considérer un modèle contraint

$$y = \mathsf{X}_c \beta_c + \epsilon,$$

où X_c est une matrice de rang $p + 1 - q$ dont l'image est inclue dans $\mathrm{Im}(\mathsf{X})$ (par exemple, pour tester l'hypothèse $\mathcal{H}_0 : \beta_j = 0$, on étudie le modèle sans la variable x_j). On définit alors la statistique de test

$$F = \frac{n-p-1}{q}\frac{\mathrm{SCR}_c - \mathrm{SCR}}{\mathrm{SCR}} = \frac{\|(\mathrm{I} - \Pi_{\mathsf{X}_c})\epsilon\|^2 - \|(\mathrm{I} - \Pi_{\mathsf{X}})\epsilon\|^2}{q\hat{\sigma}^2} = \frac{\|(\Pi_{\mathsf{X}} - \Pi_{\mathsf{X}_c})\epsilon\|^2}{q\hat{\sigma}^2}$$

où SCR_c désigne la somme des carrés résiduelle dans le modèle contraint. Si l'hypothèse $\mathcal{H}_0$ est vraie, F suit une loi de Fisher $\mathcal{F}(q, n-p-1)$. On obtient la même statistique de test par la proposition 3.9.

7.5 Prediction

On observe un nouveau jeu de variables $x_{1,n+1}, ..., x_{p,n+1}$ et on cherche à prédire la valeur y_{n+1} correspondante. On note $\mathsf{X}_{n+1} = (1, x_{1,n+1}, ..., x_{p,n+1})$. Sous l'hypothèse de normalité (qui est essentielle ici), la prédiction $\hat{y}_{n+1} = \mathsf{X}_{n+1}\hat{\beta}$ suit une loi normale $\mathcal{N}(\mathsf{X}_{n+1}\beta, \sigma^2 \mathsf{X}_{n+1}(\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}_{n+1}^\top)$ et est indépendante de $y_{n+1} = \mathsf{X}_{n+1}\beta + \epsilon_{n+1}$. On montre alors facilement que

$$\frac{\hat{y}_{n+1} - y_{n+1}}{\hat{\sigma}\sqrt{1 + \mathsf{X}_{n+1}(\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}_{n+1}^\top}} \sim \mathcal{T}(n-p-1),$$

ce qui permet de construire un intervalle de prédiction, qui n'est valable que sous l'hypothèse de normalité.

Remarque 7.13. *La longueur de l'intervalle de prédiction ne tend pas vers zéro quand n tend vers l'infini, et ce même si les observations nous permettaient d'estimer parfaitement β et σ^2 asymptotiquement. C'est dû au fait que l'intervalle de prédiction sur une nouvelle donnée y_{n+1} tient compte du bruit ϵ_{n+1} qui est indépendant du passé et donc pour lequel nous n'avons aucun moyen de prévoir les valeurs futures.*