Sorbonne Université Année 2020/2021

L3 Deuxième semestre

Statistique inférentielle

Antoine Godichon-Baggioni

Ce cours est essentiellement basé sur le cours de A. Guyader http://www.lpsm.paris/pageperso/guyader/files/teaching/StatMath.pdf

Table des matières

Introduction

Le cadre

On s'intéresse à une population de N individus, et plus particulièrement à une caractéristique a priori inconnue de cette population. Par exemple, on peut s'intéresser à la taille moyenne d'une population, savoir si le sexe des individus a une influence sur cette taille... Pour cela, on peut soit :

- effectuer un recensement, i.e mesurer chaque individu sur l'ensemble de la population. Cependant, cette méthode peut s'avérer très exhaustive, notamment si on a affaire à un grand nombre d'individus.
- procéder à un échantillonnage, i.e on réalise l'étude sur une partie de la population seulement.

On s'intéresse ici aux méthodes basées sur l'échantillonnage. En effet, elle représente généralement un coût beaucoup plus réduit que le recensement, l'étude est plus rapide et les erreurs d'observations sont plus réduites. Cependant, on n'étudie qu'une partie de la population, et si l'échantillon étudié n'est pas représentatif du reste de la population, l'extrapolation des résultats à la population globale risque d'être erronée. De plus les résultats obtenus peuvent varier d'un échantillon à l'autre.

Statistique descriptive Vs statistique inférentielle : La statistique descriptive n'a pas vocation à s'intéresser à la population globale mais juste à la sous-population formée par l'échantillon. Elle a pour but de résumer, via des tableaux, des indicateurs de position (moyenne, médiane,...) et de dispersion (variance, écart-type, écart inter-quartile...) et des graphiques, par exemple, l'information contenue par cet échantillon. La différence fondamentale entre statistique descriptive et statistique inférentielle est que l'on va supposer que les données étudiées sont des réalisations de variables aléatoires (modélisation) et s'appuyer sur la théorie des probabilités pour pouvoir extrapoler sur la population globale.

Fluctuation d'échantillonnage : Si on étudie un même caractère sur plusieurs échantillons (éventuellement de même taille) d'une même population, on peut observer que les résultats ne sont pas identiques selon les échantillons : ce phénomène s'appelle la fluctuation d'échantillonnage. Donc, les conclusions que l'on pourra tirer à l'aide d'un échantillon seront plus ou moins justes et varieront d'un échantillon à l'autre.

Prenons l'exemple du lancer d'une pièce. On va effectuer 5 lancers, et compter le nombre de fois où la pièce retombe sur "Pile". Ceci représente notre premier échantillon. On réitère notre expérience

	Echantillon 1	Echantillon 2	Echantillon 3	Echantillon 4
Nombre de "Pile"	1	3	2	3

TABLE 1 –

3 fois (ce qui représente en tous 4 échantillons) et on résume les résultats dans le tableau 1.

Objectifs du cours

On reprend l'exemple du lancer de pièce. On souhaite vérifier que la fameuse pièce n'est pas truquée. Pour cela on va effectuer n lancers, et on obtient ainsi des réalisations $x_1, \dots, x_n$, où x_i vaut 1 si le i-ème lancer vaut "Pile" et 0 sinon. On peut donc voir $x_1, \dots, x_n$ comme des réalisations des variables aléatoires indépendantes $X_1, \dots, X_n$ suivant une loi de Bernoulli de paramètre θ, et on considère donc que la pièce est équilibrée (non truquée) si $\theta = 0.5$. L'objectif de ce cours est donc :

- de faire quelques rappels de probabilités pour étudier au mieux les variables aléatoires étudiées et introduire les notions de convergence de suites de variables aléatoires. Dans l'exemple du lancer de pièce, on a compté le nombres de piles, i.e on a regardé $\sum_{i=1}^{5} x_i$ qui est donc la réalisation de la variable aléatoire $\sum_{i=1}^{5} X_i$. Quelle est la loi de cette variable aléatoire ? Que se passe-t-il si on effectue plus (une infinité) de lancers ?
- d'introduire une ou plusieurs méthodes d'estimations des paramètres qui nous intéressent : comment estimer le paramètre θ de la loi de Bernoulli qui "régit" ma pièce ?
- d'apprendre à comparer ces différents estimateurs pour savoir lequel choisir.
- de construire des intervalles de confiances pour les différents paramètres : avec un certain risque, dire que θ appartient à un certain intervalle.
- de pouvoir répondre, avec un certain risque, si la pièce est truquée ou non.

Chapitre 1

Convergence de suites de variables aléatoires

1.1 Rappels de probabilités

Dans ce qui suit, on considère un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et on considère une variable aléatoire $X : \Omega \longrightarrow \mathbb{R}$. On rappelle que la fonction de répartition $F_X : \mathbb{R} \longrightarrow [0,1]$ de X est définie pour tout $x \in \mathbb{R}$ par

$$F_X(x) = \mathbb{P}\left[X \leq x\right].$$

Rappelons que cette fonction est croissante, continue à droite et

$$\lim_{x \to -\infty} F_X(x) = 0, \qquad \text{et} \qquad \lim_{x \to +\infty} F_X(x) = 1.$$

Exemples :

1. Si X suit une loi de Bernoulli de paramètre p,

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1-p & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

2. Si X suit une loi uniforme sur $[0,1]$,

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

3. Si X suit une loi normale centrée réduite,

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt.$$

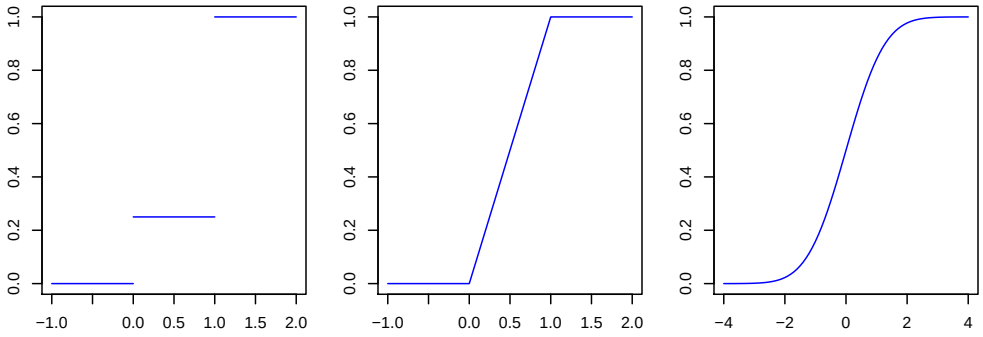

FIGURE 1.1 – Fonctions de répartition d'une loi de Bernoulli de paramètre $p = 0.75$ (à gauche), d'une loi uniforme (au centre) et d'une loi normale centrée réduite (à droite).

1.2 Modes de convergence

Dans ce qui suit, on s'intéresse aux différents modes de convergence d'une suite de variables aléatoires (X_n). Plus précisément, on s'intéresse aux convergences en loi, en probabilité, presque sûre et en moyenne quadratique.

1.2.1 Convergence en loi

Definition 1.2.1 (Convergence en loi). *On dit que la suite (X_n) converge en loi vers une variable aléatoire X et on note*

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$$

si pour toute fonction continue et bornée φ, on a

$$\mathbb{E}\left[\varphi\left(X_n\right)\right] \xrightarrow[n\to+\infty]{} \mathbb{E}\left[\varphi\left(X\right)\right].$$

De manière équivalente, (X_n) converge en loi vers X si pour toute fonction uniformément continue et bornée φ,

$$\mathbb{E}\left[\varphi\left(X_n\right)\right] \xrightarrow[n\to+\infty]{} \mathbb{E}\left[\varphi\left(X\right)\right].$$

Exemple 1 : On considère une suite de variables aléatoires (X_n), où pour tout $n \geq 1$, $X_n \sim \mathcal{B}\left(p + \frac{1-p}{n}\right)$, alors

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X \sim \mathcal{B}\left(p\right).$$

En effet, pour toute fonction continue et bornée φ, on a

$$\mathbb{E}\left[\varphi\left(X_n\right)\right] = \varphi\left(1\right)\left(p + \frac{1-p}{n}\right) + \varphi(0)\frac{n-1}{n}(1-p) \xrightarrow[n\to+\infty]{} p\varphi(1) + (1-p)\varphi(0) = \mathbb{E}\left[\varphi(X)\right].$$

Exemple 2 : On considère la suite (X_n) avec pour tout n, $X_n = -X$ où $X \sim \mathcal{N}\left(0,1\right)$. Comme la loi normale est symétrique, on a $-X \sim \mathcal{N}\left(0,1\right)$ et donc

$$X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X.$$

Exemple 3 : Pour tout $n \geq 1$, on considère $X_n \sim \mathcal{N}\left(0, \sigma^2 + \frac{1}{n}\right)$, on a pour toute fonction continue bornée φ, par convergence dominée

$$\mathbb{E}\left[\varphi\left(X_n\right)\right] = \int_{-\infty}^{\infty} \varphi(x) \frac{1}{\sqrt{2\pi\left(\sigma^2 + \frac{1}{n}\right)}} \exp\left(-\frac{x^2}{2\left(\sigma^2 + \frac{1}{n}\right)}\right) dx \xrightarrow[n\to+\infty]{} \int_{-\infty}^{\infty} \varphi(x) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = \mathbb{E}\left[\varphi(X)\right],$$

où $X \sim \mathcal{N}\left(0, \sigma^2\right)$ et donc X_n converge en loi vers X.

Proposition 1.2.1. *La suite de variables aléatoires (X_n) converge en loi vers X si et seulement si en tout point de continuité x de la fonction de répartition F_X de X,*

$$F_{X_n}(x) \xrightarrow[n\to+\infty]{} F_X(x).$$

Exemple 1 : On considère pour tout $n \geq 1$, $X_n \sim \mathcal{B}\left(p + \frac{1-p}{n}\right)$. On a alors pour tout $x \in \mathbb{R}$,

$$F_{X_n}(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{n-1}{n}(1-p) & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

On a donc $F_{X_n}(x)$ qui converge vers $F_X(x)$ pour tout $x \notin \{0,1\}$, où $X \sim \mathcal{B}(p)$.

Exemple 2 : Pour tout $n \geq 1$, on considère $X_n = -X$, où $X \sim \mathcal{N}\left(0,1\right)$, alors pour tout $x \in \mathbb{R}$, par symétrie de la loi normale,

$$F_{X_n}(x) = \mathbb{P}\left[X_n \leq x\right] = \mathbb{P}\left[-X \leq x\right] = \mathbb{P}\left[X \geq -x\right] = \mathbb{P}\left[X \leq x\right] = F_X(x).$$

Exemple 3 : Pour tout n, $X_n \sim \mathcal{N}\left(0, \sigma^2 + \frac{1}{n}\right)$. On a alors pour tout x,

$$F_{X_n}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\left(\sigma^2 + \frac{1}{n}\right)}} \exp\left(-\frac{t^2}{2\left(\sigma^2 + \frac{1}{n}\right)}\right) dt.$$

Par convergence dominée,

$$F_{X_n}(x) \xrightarrow[n\to+\infty]{} \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2}} dt = F_X(x)$$

où $X \sim \mathcal{N}(0,1)$. On a donc la convergence en loi de X_n vers X.

Corollaire 1.2.1. *Supposons que les variables aléatoires X_n et X sont absolument continues de densités f_n et f. Alors, X_n converge en loi vers X si et seulement si pour tout $x \in \mathbb{R}$,*

$$f_n(x) \xrightarrow[n\to+\infty]{} f(x).$$

Si on reprend l'exemple 3, on obtient directement pour tout $x \in \mathbb{R}$,

$$f_n(x) \xrightarrow[n\to+\infty]{} f(x).$$

Enfin, rappelons qu'une variable aléatoire est caractérisée par sa fonction caractéristique, définie pour tout $t \in \mathbb{R}$ par $\Phi_X(t) = \mathbb{E}\left[e^{itX}\right]$. On peut alors réécrire la Proposition 1.2.1 comme :

Proposition 1.2.2. *La suite de variables aléatoires (X_n) converge en loi vers X si et seulement si pour tout $t \in \mathbb{R}$,*

$$\Phi_{X_n}(t) \xrightarrow[n\to+\infty]{} \Phi_X(t).$$

Passer par les fonctions caractéristiques peut représenter un grand intérêt lorsque l'on doit traiter des sommes de variables aléatoires indépendantes. En effet, rappelons que pour toutes variables aléatoires indépendantes X, Y,

$$\Phi_{X+Y}(t) = \Phi_X(t)\Phi_Y(t).$$

Exemple : On considère l'exemple introductif du lancer d'une pièce de monnaie, et on considère les variables aléatoires X_i prenant comme valeur 1 si le i-ème lancer vaut "Pile" et 0 sinon. On a alors pour tout i, $X_i \sim \mathcal{B}(\theta)$ et un estimateur naturel de θ est $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. On considère les lancers comme étant indépendants et on a

$$\Phi_{\hat{\theta}_n}(t) = \Phi_{\frac{X_1}{n}}(t) \dots \Phi_{\frac{X_n}{n}}(t) = \left(\Phi_{\frac{X_1}{n}}(t)\right)^n = \left(1 - \theta + \theta e^{i\frac{t}{n}}\right)^n \xrightarrow[n\to+\infty]{} e^{it\theta}$$

et on a donc la convergence en loi de $\hat{\theta}_n$ vers θ.

1.2.2 Convergence en probabilité

Definition 1.2.2 (Convergence en probabilité). *On dit que la suite (X_n) converge en probabilité vers une variable aléatoire X et on note*

$$X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X$$

si pour tout $\epsilon > 0$,

$$\mathbb{P}\left[|X_n - X| > \epsilon\right] \xrightarrow[n\to+\infty]{} 0$$

Exemple 1 : On considère pour tout $n \geq 1$, $X_n \sim \mathcal{B}\left(\frac{p}{n}\right)$ avec $p \in (0,1)$. Alors X_n converge en probabilité vers 0. En effet pour tout $\epsilon \in (0,1)$,

$$\mathbb{P}\left[|X_n - 0| > \epsilon\right] = \mathbb{P}\left[X_n > \epsilon\right] = \mathbb{P}\left[X_n = 1\right] = \frac{p}{n} \xrightarrow[n\to+\infty]{} 0.$$

Exemple 2 : On considère des variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$ avec $X_1 \sim \mathcal{U}([0,\theta])$, et on considère l'estimateur $X_{(n)} = \max_{i=1,\ldots,n} X_i$. Alors $X_{(n)}$ converge presque sûrement vers θ. En effet, comme $X_{(n)} \leq \theta$, pour tout $\epsilon > 0$ on a

$$\mathbb{P}\left[\left|X_{(n)} - \theta\right| \geq \epsilon\right] = \mathbb{P}\left[\theta - X_{(n)} \geq \epsilon\right] = \mathbb{P}\left[X_{(n)} \leq \theta - \epsilon\right] = \mathbb{P}\left[\forall i, X_i \leq \theta - \epsilon\right]$$

A noter que cette probabilité est nulle si $\epsilon \geq \theta$, et on va donc prendre $\epsilon \in (0,\theta)$. De plus, comme les variables X_i sont indépendantes, on a

$$\mathbb{P}\left[\left|X_{(n)} - \theta\right| \geq \epsilon\right] = \prod_{i=1}^{n} \mathbb{P}\left[X_i \leq \theta - \epsilon\right] = \left(\mathbb{P}\left[X_1 \leq \theta - \epsilon\right]\right)^n = \left(\frac{\theta - \epsilon}{\theta}\right)^n \xrightarrow[n\to+\infty]{} 0.$$

La proposition suivante donne le lien entre la convergence en loi et la convergence en probabilité.

Proposition 1.2.3. *Soit (X_n) une suite de variables aléatoires. Si (X_n) converge en probabilité vers X, alors (X_n) converge en loi vers X.*

Démonstration. Soit φ une fonction uniformément continue, et soit $M > 0$ telle que pour tout x, $|\varphi(x)| \leq M$. Soit $\epsilon > 0$, par continuité uniforme de φ, il existe $\eta > 0$ tel que pour tout x, y avec $|x - y| \leq \eta$, on ait $|\varphi(x) - \varphi(y)| \leq \epsilon$. On a alors

$$\begin{aligned}
\left|\mathbb{E}\left[\varphi\left(X_n\right)\right] - \mathbb{E}\left[\varphi(X)\right]\right| &\leq \mathbb{E}\left[\left|\varphi\left(X_n\right) - \varphi(X)\right|\right] \\
&= \mathbb{E}\left[\left|\varphi\left(X_n\right) - \varphi(X)\right| \mathbb{1}_{\{|X_n - X| < \eta\}}\right] + \mathbb{E}\left[\left|\varphi\left(X_n\right) - \varphi(X)\right| \mathbb{1}_{\{|X_n - X| \geq \eta\}}\right] \\
&\leq \epsilon + 2 \sup_x \varphi(x) \mathbb{E}\left[\mathbb{1}_{\{|X_n - X| \geq \eta\}}\right] \\
&\leq \epsilon + 2M\mathbb{P}\left[|X_n - X| \geq \eta\right].
\end{aligned}$$

Comme (X_n) converge en probabilité vers X, il existe un rang n_ϵ tel que pour tout $n \geq n_\epsilon$,

$$\left|\mathbb{E}\left[\varphi\left(X_n\right)\right] - \mathbb{E}[\varphi(X)]\right| \leq 2\epsilon,$$

et on obtient donc la convergence en loi de (X_n) vers X. □

Remarque : Attention la réciproque de la proposition précédente est généralement fausse.
Contre-exemple : Reprenons l'exemple naïf précédent, où X suit une loi normale centrée réduite et pour tout n, $X_n = -X$. On a vu que (X_n) converge en loi vers X (par symétrie), mais il est évident que X_n ne converge pas en probabilité vers X. En effet, prenons $\epsilon = 3.92$, on a alors

$$\mathbb{P}\left[|X_n - X| > \epsilon\right] = \mathbb{P}\left[|2X| > \epsilon\right] = 0.05$$

et ne converge donc pas vers 0.

Proposition 1.2.4. *Soit (X_n) une suite de variables aléatoires qui converge en loi vers une constante c. Alors (X_n) converge en probabilité vers c.*

Démonstration. Pour tout $\epsilon > 0$, on a

$$\begin{aligned}\mathbb{P}\left[|X_n - c| > \epsilon\right] &= \mathbb{P}\left[X_n - c > \epsilon\right] + \mathbb{P}\left[c - X_n > \epsilon\right] \\ &= 1 - F_{X_n}\left[\epsilon + c\right] + F_{X_n}\left[X_n \leq c - \epsilon\right] - \mathbb{P}\left[X_n = c - \epsilon\right].\end{aligned}$$

Comme pour tout $x \in \mathbb{R}$, on a $F_c(x) = \mathbf{1}_{x \geq c}$, on a que $x = c$ est le seul point de discontinuité de F_c et comme $\epsilon > 0$,

$$F_{X_n}(\epsilon + c) \xrightarrow[n \to +\infty]{} F_c(\epsilon + c) = 1 \qquad \text{et} \qquad F_{X_n}(c - \epsilon) \xrightarrow[n \to +\infty]{} F_c(\epsilon - c) = 0.$$

Enfin, comme une fonction de répartition est continue à droite, on a

$$\lim_{n \to +\infty} \mathbb{P}\left[X_n = c - \epsilon\right] = \lim_{n \to +\infty} \lim_{\substack{h \to 0 \\ >}} F_{X_n}\left(c - \epsilon + h\right) = \lim_{\substack{h \to 0 \\ >}} \lim_{n \to +\infty} F_{X_n}\left(c - \epsilon + h\right) = \lim_{\substack{h \to 0 \\ >}} F_c(c - \epsilon + h) = 0.$$

□

1.2.3 Convergence presque sûre

Definition 1.2.3 (Convergence presque sûre). *On dit que la suite (X_n) converge presque sûrement vers X et on note*

$$X_n \xrightarrow[n \to +\infty]{p.s} X$$

si

$$\mathbb{P}\left[\lim_{n \to +\infty} X_n = X\right] = \mathbb{P}\left[\left\{\omega \in \Omega; \quad \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right] = 1.$$

La proposition suivante donne le lien entre la convergence presque sûre et la convergence en probabilité.

Proposition 1.2.5. *Soit (X_n) une suite de variables aléatoires. Si (X_n) converge presque sûrement vers X, alors (X_n) converge en probabilité vers X.*

Attention, la réciproque est fausse.

Contre-exemple. On considère une variable X suivant une loi uniforme sur $[0,1]$ et la suite d'évènements (A_n) définie par $A_1 = \{X \in [0,1]\}$, $A_2 = \{X \in [0,1/2]\}$, $A_3 = \{X \in [1/2,1]\}$, $A_4 = \{X \in [0,1/4]\}$,.... et on considère la suite de variable aléatoire (Y_n) définie pour tout $n \geq 1$ par $Y_n = \mathbf{1}_{A_n}$. On peut montrer que

$$\mathbb{P}\left[A_n\right] = \left(\frac{1}{2}\right)^{[\log_2(n)]}$$

où $[.]$ est la fonction partie entière, et on a donc Y_n qui converge en probabilité vers 0. Cependant, pour tout $\omega \in [0,1]$, pour tout $n \geq 1$, il existe n_ω tel que $X(\omega) \in A_{n_\omega}$, i.e

$$Y_n(\omega) = \mathbf{1}_{X(\omega)\in A_n} \nrightarrow 0$$

et on n'a donc pas convergence presque sûre.

Cependant, grâce au lemme de Borel-Cantelli, si pour tout $\epsilon > 0$,

$$\sum_{n=1}^{+\infty} \mathbb{P}\left[|X_n - X| \geq \epsilon\right] < +\infty,$$

alors (X_n) converge presque sûrement vers X.

Exemple : Soit (X_n) une suite de variables aléatoires telle que $X_n \sim \mathcal{B}\left(\frac{1}{n^2}\right)$. Alors

$$X_n \xrightarrow[n\to+\infty]{p.s} 0.$$

En effet pour tout $\epsilon \in (0,1)$, on a

$$\mathbb{P}\left[X_n \geq \epsilon\right] = \mathbb{P}\left[X_n = 1\right] = \frac{1}{n^2} \qquad \text{et} \qquad \sum_{n\geq 1} \mathbb{P}\left[X_n \geq \epsilon\right] = \sum_{n\geq 1} \frac{1}{n^2} < +\infty.$$

1.2.4 Convergence en moyenne quadratique

Definition 1.2.4 (Convergence en moyenne quadratique). *Soit (X_n) une suite de variables aléatoires de carré intégrable. On dit que (X_n) converge en moyenne quadratique vers X et on note*

$$X_n \xrightarrow[n\to+\infty]{L^2} X$$

si

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

Exemple : Reprenons l'exemple du lancer de pièce, i.e où $X_1, ..., X_n$ sont i.i.d et pour tout i, $X_i \sim \mathcal{B}(\theta)$. On a $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et par linéarité de l'espérance

$$\mathbb{E}\left[\hat{\theta}_n\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] = \theta.$$

On a alors

$$\mathbb{E}\left[\left|\hat{\theta}_n - \theta\right|^2\right] = \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2\right] = \mathbb{V}\left[\hat{\theta}_n\right].$$

Par indépendance, on a

$$\mathbb{V}\left[\hat{\theta}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}\left[X_i\right] = \frac{\theta(1-\theta)}{n},$$

et on obtient donc la convergence en moyenne quadratique de $\hat{\theta}_n$ vers θ .
La proposition suivante donne le lien entre convergence en probabilité et convergence en moyenne quadratique.

Proposition 1.2.6. *Soit (X_n) une suite de variables aléatoires. Si (X_n) converge en moyenne quadratique vers X, alors (X_n) converge en probabilité vers X.*

Démonstration. Voir TD. □

Attention la réciproque est fausse.

Contre-exemple : On considère une suite de variables aléatoires discrètes (X_n) avec pour tout n,

$$\mathbb{P}[X_n = 0] = \frac{n-1}{n}, \qquad \mathbb{P}[X_n = n] = \frac{1}{n}.$$

Soit $\epsilon > 0$, pour tout $n \geq \epsilon^{-1}$ (et $n \geq \epsilon$), on a

$$\mathbb{P}[X_n \geq \epsilon] = \mathbb{P}[X_n = n] = \frac{1}{n},$$

et donc X_n converge en probabilité vers 0. Cependant,

$$\mathbb{E}\left[X_n^2\right] = n \xrightarrow[n\to+\infty]{} +\infty.$$

1.2.5 Résumé

Le tableau suivant résume les liens entre les différents types de convergence.

$$X_n \xrightarrow[n\to+\infty]{p.s} X \implies X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$$

$$X_n \xrightarrow[n\to+\infty]{L^2} X \implies X_n \xrightarrow[n\to+\infty]{\mathbb{P}} X \implies X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$$

Remarque : Il n'y a pas d'implication générale entre la convergence en moyenne quadratique et la convergence presque sûre.

Contre-exemple 1 : Soit X une variable aléatoire suivant une loi uniforme sur $[0,1]$ et la suite d'évènements (A_n) définie par $A_1 = \{X \in [0,1]\}$, $A_2 = \{X \in [0,1/2]\}$, $A_3 = \{X \in [1/2,1]\}$, $A_4 = \{X \in [0,1/4]\}$,.... et on considère la suite de variable aléatoire (Y_n) définie pour tout $n \geq 1$ par $Y_n = \mathbf{1}_{A_n}$. On rappelle que

$$\mathbb{P}[A_n] = \left(\frac{1}{2}\right)^{[\log_2(n)]}$$

et on a donc

$$\mathbb{E}\left[Y_n^2\right] = \mathbb{E}[\mathbf{1}_{A_n}] = \left(\frac{1}{2}\right)^{[\log_2(n)]} \xrightarrow[n\to+\infty]{} 0$$

et on a donc convergence en moyenne quadratique. Cependant, on a vu que l'on n'a pas convergence presque sûre.

Contre-exemple 2 : Soit (X_n) une suite de variables aléatoires telle que $\mathbb{P}[X_n = 0] = 1 - \frac{1}{n^2}$ et $\mathbb{P}[X_n = n] = \frac{1}{n^2}$. Alors (X_n) converge presque sûrement vers 0 mais pas en moyenne quadratique. En effet, pour tout $\epsilon \in (0,1)$,

$$\sum_{n\geq 1} \mathbb{P}[X_n \geq \epsilon] = \sum_{n\geq 1} \frac{1}{n^2} < +\infty$$

d'où la convergence presque sûre. Cependant,

$$\mathbb{E}\left[X_n^2\right] = \frac{n^2}{n^2} = 1.$$

1.3 Quelques inégalités classiques

Lorsque l'on considère une suite de variables aléatoires (X_n) et que l'on ne connait pas leurs lois, il n'est pas aisé de démontrer leur convergence en probabilité directement. Les inégalités suivantes peuvent alors servir.

Proposition 1.3.1 (Inégalité de Markov). *Soit $c, p > 0$ et X une variable aléatoire admettant un moment d'ordre p, on a*

$$\mathbb{P}[|X| \geq c] \leq \frac{\mathbb{E}\left[|X|^p\right]}{c^p}.$$

Démonstration. On a

$$|X|^p = |X|^p \mathbb{1}_{\{|X|<c\}} + |X|^p \mathbb{1}_{\{|X|\geq c\}} \geq c^p \mathbb{1}_{\{|X|\geq c\}}$$

En passant à l'espérance, on obtient

$$\mathbb{E}\left[|X|^p\right] \geq c^p \mathbb{E}\left[\mathbb{1}_{\{|X|\geq c\}}\right] = c^p \mathbb{P}\left[|X| \geq c\right].$$

□

L'inégalité de Markov est par exemple très utile pour montrer que la convergence en moyenne quadratique implique la convergence en probabilité (voir TD).

Corollaire 1.3.1 (Inégalité de Bienaymé-Tchebychev). *Soit $c > 0$ et X une variable aléatoire admettant un moment d'ordre 2. Alors,*

$$\mathbb{P}\left[|X - \mathbb{E}\left[X\right]| \geq c\right] \leq \frac{\mathbb{V}[X]}{c^2}$$

Démonstration. C'est une application directe de l'inégalité de Markov, en considérant la variable aléatoire $Y = X - \mathbb{E}[X]$ et en prenant $p = 2$. En effet, on a alors

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq c\right] = \mathbb{P}\left[|Y| \geq c\right] \leq \frac{\mathbb{E}\left[Y^2\right]}{c^2} = \frac{\mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right]}{c^2} = \frac{\mathbb{V}[X]}{c^2}.$$

□

Exemple : On considère l'exemple introductif du lancer d'une pièce. Rappelons que l'on considère les variables aléatoires X_i prenant comme valeur 1 si le i-ème lancé vaut "Pile" et 0 sinon. Rappelons qu'un estimateur naturel du paramètre θ est $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, et comme $\mathbb{E}\left[\hat{\theta}_n\right] = \theta$ et $\mathbb{V}\left[\hat{\theta}_n\right] = \frac{\theta(1-\theta)}{n}$, on obtient pour tout $\epsilon > 0$, à l'aide de l'inégalité de Bienaymé-Tchebychev

$$\mathbb{P}\left[\left|\hat{\theta}_n - \theta\right| \geq \epsilon\right] \leq \frac{\mathbb{V}\left[\hat{\theta}_n\right]}{\epsilon^2} = \frac{\theta(1-\theta)}{\epsilon^2 n},$$

et donc la convergence en probabilité de $\hat{\theta}_n$ vers θ.

Théorème 1.3.1 (Inégalité de Cauchy-Schwarz). *Soit X, Y deux variables aléatoires admettant un moment d'ordre 2. Alors*

$$\mathbb{E}\left[|XY|\right] \leq \sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}.$$

Démonstration. On considère la fonction

$$\lambda \longmapsto \mathbb{E}\left[\left(\lambda|X| - |Y|\right)^2\right] = \lambda^2\mathbb{E}\left[X^2\right] - 2\lambda\mathbb{E}\left[|XY|\right] + \mathbb{E}\left[Y^2\right].$$

On a un polynôme positif, et son discriminant $\Delta = 4\left(\mathbb{E}\left[|XY|\right]\right)^2 - 4\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]$ est donc négatif ou nul, i.e

$$\left(\mathbb{E}\left[|XY|\right]\right)^2 \leq \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right].$$

□

L'inégalité de Cauchy Schwarz peut être particulièrement utile pour obtenir la convergence en moyenne quadratique de produits d'estimateurs.

Exemple : Soit $(X_n), (Y_n)$ deux suites d'estimateurs convergeant en moyenne quadratique vers 0. On suppose qu'ils convergent également à l'ordre 4, i.e ils admettent des moments d'ordre 4 et

$$\mathbb{E}\left[|X_n - 0|^4\right] \xrightarrow[n\to+\infty]{} 0 \qquad \text{et} \qquad \mathbb{E}\left[|Y_n - 0|^4\right] \xrightarrow[n\to+\infty]{} 0.$$

Alors X_nY_n converge en moyenne quadratique vers 0. En effet, on a

$$\mathbb{E}\left[\left(X_nY_n\right)^2\right] \leq \sqrt{\mathbb{E}\left[X_n^4\right]\mathbb{E}\left[Y_n^4\right]} \xrightarrow[n\to+\infty]{} 0,$$

i.e on a donc bien convergence en moyenne quadratique. Cependant, l'inégalité de Cauchy-Schwarz est assez grossière il est parfois nécessaire d'avoir des inégalités un peu moins grossière, ce que nous donne l'inégalité de Hölder.

Théorème 1.3.2 (Inégalité de Hölder). *Soit X, Y deux variables aléatoires, et $p, q > 1$ tel que $\frac{1}{p} + \frac{1}{q} = 1$. Si X admet un moment d'ordre p et si Y admet un moment d'ordre q, alors*

$$\mathbb{E}\left[|XY|\right] \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}}\left(\mathbb{E}\left[|Y|^q\right]\right)^{\frac{1}{q}}.$$

Démonstration. Remarquons que pour tout $a, b \geq 0$, on a

$$ab \leq \frac{1}{p}a^p + \frac{1}{q}b^q.$$

En effet, si on considère la fonction $g_b : a \mapsto \frac{1}{p}a^p + \frac{1}{q}b^q - ab$, on a

$$g_b'(a) = a^{p-1} - b$$

et le minimum atteint en $a^* = b^{\frac{1}{p-1}}$. De plus, comme $q = \frac{p}{p-1}$,

$$g_b\left(a^*\right) = \frac{1}{p}b^{\frac{p}{p-1}} + \frac{1}{q}b^{\frac{p}{p-1}} - b^{\frac{p}{p-1}} = 0.$$

En prenant

$$a = \frac{|X|}{\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}} \qquad \text{et} \qquad b = \frac{|Y|}{\left(\mathbb{E}\left[|Y|^q\right]\right)^{1/q}},$$

on obtient

$$\frac{|XY|}{\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}\left(\mathbb{E}\left[|Y|^q\right]\right)^{1/q}} \leq \frac{|X|^p}{p\mathbb{E}\left[|X|^p\right]} + \frac{|Y|^q}{q\mathbb{E}\left[|Y|^q\right]}$$

et en passant a l'espérance

$$\frac{\mathbb{E}\left[|XY|\right]}{\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p}\left(\mathbb{E}\left[|Y|^q\right]\right)^{1/q}} \leq \frac{1}{p} + \frac{1}{q} = 1.$$

□

Exemple : Si X admet un moment d'ordre r, alors pour tout $r' \in (0, r)$,

$$\mathbb{E}\left[|X|^{r'}\right] \leq \left(\mathbb{E}\left[|X|^r\right]\right)^{\frac{r'}{r}}.$$

Exemple : Comme dit précédemment, l'inégalité de Hölder permet un certain "raffinement" dans les calculs. Si on reprend l'exemple précédent mais en supposant cette fois que X_n converge à l'ordre r avec $2 < r < 4$, les calculs précédents ne sont plus valables. Cependant, si on suppose que Y_n converge à l'ordre $\frac{2r}{r-2}$, on a à l'aide de l'inégalité de Hölder (en prenant $p = r/2$ et $q = \frac{r}{r-2}$)

$$\mathbb{E}\left[(X_nY_n)^2\right] \leq \left(\mathbb{E}\left[|X_n|^r\right]\right)^{\frac{2}{r}}\left(\mathbb{E}\left[|Y_n|^{\frac{2r}{r-2}}\right]\right)^{\frac{r-2}{r}}.$$

et ce terme converge donc vers 0, i.e on a bien convergence en moyenne quadratique.

1.4 Théorèmes asymptotiques

A travers l'exemple précédent, on a vu comment obtenir la convergence en probabilité d'une suite de variables aléatoires. On rappelle ici les théorèmes classiques qui permettent d'obtenir rapidement ce type de (ou de meilleurs) résultats. Dans ce qui suit, on considère une suite de variables aléatoires indépendantes et identiquement distribuées (X_i) et on note

$$S_n := X_1 + \ldots + X_n.$$

Théorème 1.4.1 (**Loi faible des Grands Nombres (LGN)**). *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées admettant une moyenne $m = \mathbb{E}[X_1]$, alors*

$$\frac{S_n}{n} \xrightarrow[n\to+\infty]{\mathbb{P}} m.$$

A noter que pour la loi faible des grands nombres, il est souvent noté comme hypothèse $\mathbb{E}[|X_1|] < +\infty$. En réalité, les deux hypothèses sont équivalentes car on ne peut parler d'espérance de X_1 que si X_1 admet un moment d'ordre 1. De plus, il est fréquent de voir la loi faible des grands nombres avec des hypothèses sur l'existence de la variance de X_1 (permettant ainsi d'utiliser l'inégalité de Bienaymé-Tchebychev pour démontrer le résultat, à l'image de ce qui a été fait dans les exemples précédents). Cependant, cette hypothèse est inutile, notamment si on voit la loi faible des grands nombres comme un corollaire de la loi forte des grands nombres ci-dessous :

Théorème 1.4.2 (**Loi Forte des Grands Nombres (LFGN)**). *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées admettant une moyenne $m = \mathbb{E}[X_1]$, alors*

$$\frac{S_n}{n} \xrightarrow[n\to+\infty]{p.s} m.$$

Exemple : Reprenons l'exemple du pile ou face. On a des variables aléatoires i.i.d et d'espérance $\mathbb{E}[X_1] = \theta$. En appliquant la loi forte des grands nombres, on obtient

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{p.s} \theta,$$

et on a en particulier la convergence en probabilité.

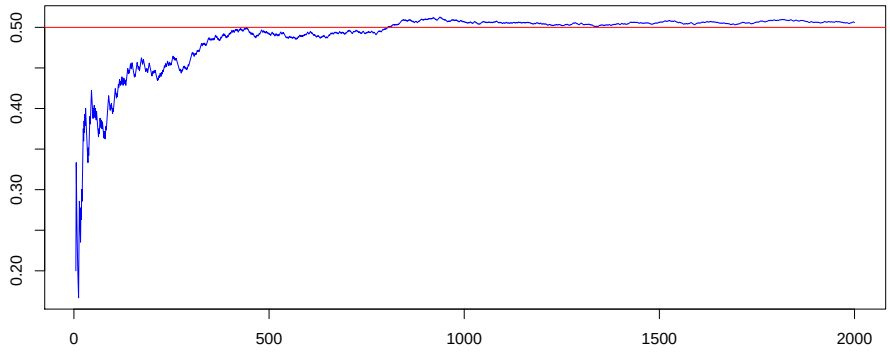

FIGURE 1.2 – Evolution en fonction de n d'une réalisation de $\hat{\theta}_n$ avec $\theta = 0.5$

La figure 1.2 laisse penser que l'estimateur $\hat{\theta}_n$ converge bien vers θ, et qu'il converge même rapidement. Les loi fortes des grands nombres, bien que très utiles pour obtenir "facilement" la convergence d'estimateurs, ne donnent malheureusement aucune indication sur la vitesse de convergence, contrairement au Théorème Central Limite :

Théorème 1.4.3 (Théorème Central Limite). *Soit (X_n) une suite de variables aléatoires indépendantes et identiquement distribuées de moyenne $m = \mathbb{E}[X_1]$ et admettant une variance $\sigma^2 = \mathbb{V}[X_1]$, alors*

$$\sqrt{n}\left(\frac{S_n}{n} - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

ce qui peut également s'écrire

$$\frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - m\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, 1),$$

Démonstration. A noter qu'en notant $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, on peut réécrire

$$Z_n := \frac{\sqrt{n}}{\sigma}\left(\frac{S_n}{n} - m\right) = \sqrt{n}\frac{\overline{X}_n - m}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \underbrace{\frac{X_i - m}{\sigma}}_{=:Y_i}.$$

A noter que les Y_i sont i.i.d et en regardant la fonction caractéristique, on a pour tout $t \in \mathbb{R}$,

$$\Phi_{Z_n}(t) = \mathbb{E}\left[e^{it\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{i\frac{t}{\sqrt{n}}Y_i}\right] = \left(\mathbb{E}\left[e^{i\frac{t}{\sqrt{n}}Y_1}\right]\right)^n.$$

On considère la fonction $\varphi : \mathbb{R} \longrightarrow \mathbb{C}$ définie pour tout $u \in \mathbb{R}$ par $\varphi(u) = \mathbb{E}\left[e^{iuY_1}\right]$. Comme Y_1 admet un moment d'ordre 2, cette fonction est deux fois dérivable et

$$\varphi'(u) = i\mathbb{E}\left[Y_1 e^{iuY_1}\right] \qquad \text{et} \qquad \varphi''(u) = -\mathbb{E}\left[Y_1^2 e^{iuY_1}\right].$$

A l'aide d'un développement de Taylor, il existe une fonction g continue en 0 et telle que $g(0) = 0$ vérifiant

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}u^2\left(\varphi''(0) + g(u)\right)$$

et on a donc

$$\varphi(u) = 1 + iu\mathbb{E}\left[Y_1\right] - \frac{1}{2}u^2\left(\mathbb{E}\left[Y_1^2\right] + g(u)\right) = 1 - \frac{1}{2}u^2\left(1 + g(u)\right).$$

Pour tout t, on a donc

$$\varphi\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n}\left(1 + \epsilon\left(\frac{t}{\sqrt{n}}\right)\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t}{2n}\right).$$

On a donc

$$\Phi_{Z_n}(t) = \varphi\left(\frac{t}{\sqrt{n}}\right)^n = \exp\left(n\log\varphi\left(\frac{t}{\sqrt{n}}\right)\right) \xrightarrow[n\to+\infty]{} \exp\left(-\frac{t^2}{2}\right)$$

et on retrouve ainsi la fonction caractéristique d'une loi normale centrée réduite, ce qui conclut la preuve.

□

Remarque : Le TCL est également appelé Théorème de la Limite Centrale (TLC).

Notons que dans l'exemple de la pièce équilibrée, on a $\mathbb{E}\left[X_1\right] = 1/2$, $\mathbb{V}\left[X_1\right] = 1/4$, et le TCL s'écrit alors

$$\sqrt{n}\left(\hat{\theta}_n - \frac{1}{2}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{4}\right),$$

ce que l'on peut également écrire

$$P_n := \sqrt{n}\left(2\hat{\theta}_n - 1\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, 1\right).$$

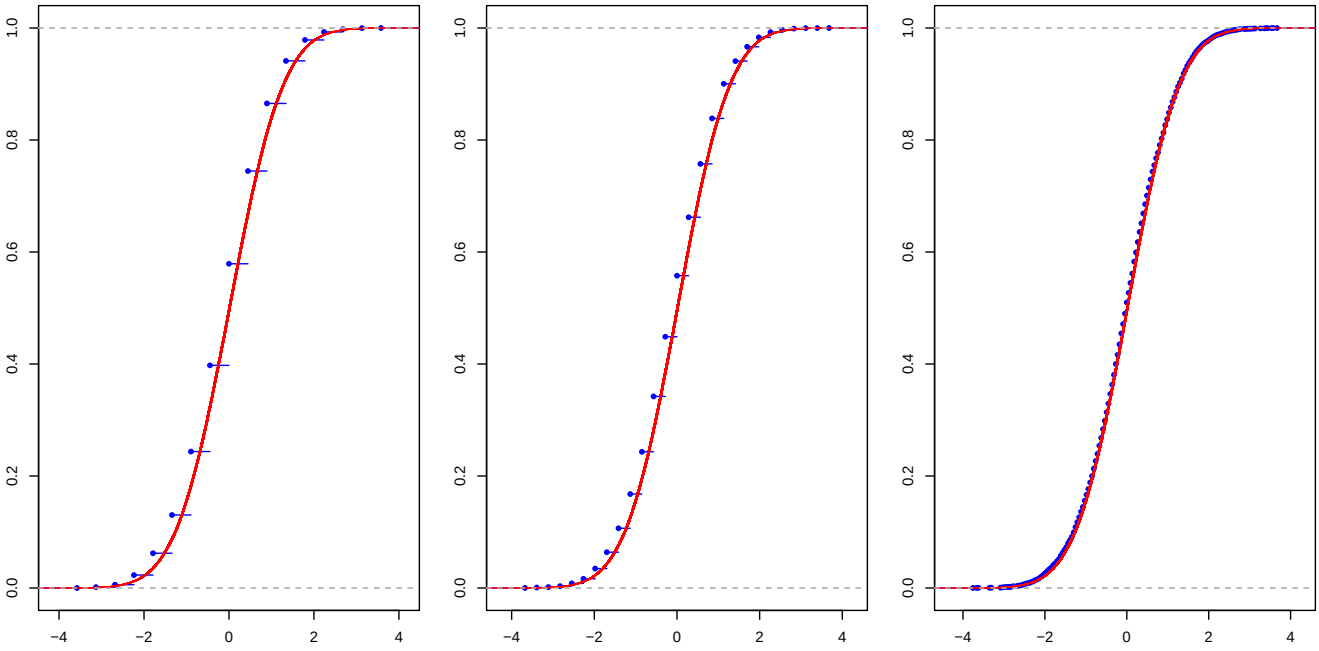

FIGURE 1.3 – Fonctions de répartition de P_n (en bleu) et d'une loi normale centrée réduite (en rouge) pour $n = 20$ (à gauche) $n = 50$ (au centre) et $n = 2000$ (à droite).

1.5 Opérations sur les limites

Si on reprend l'exemple du lancer de pièce, mais cette fois ci sans savoir si la pièce est équilibrée, i.e le paramètre θ est inconnu, le TLC nous donne alors

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

que l'on peut également écrire comme

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

Celui-ci est alors inexploitable en l'état (pour construire des intervalles de confiance par exemple) car le paramètre θ est inconnu, et on ne peut donc pas calculer $\sqrt{\theta(1-\theta)}$. Cependant, un estimateur naturel serait $\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}$, et la proposition suivante peut aider à montrer sa convergence.

Théorème 1.5.1 (Théorème de continuité). *Soit (X_n) une suite de variables aléatoires et c une constante. Soit g une fonction une fonction continue en c, alors la suite $g(X_n)$ hérite du mode de convergence de la suite (X_n), i.e*

1. *Si (X_n) converge presque sûrement vers c, alors $g(X_n)$ converge presque sûrement vers $g(c)$.*
2. *Si (X_n) converge en probabilité vers c, alors $g(X_n)$ converge en probabilité vers $g(c)$.*
3. *Si (X_n) converge en loi vers c, alors $g(X_n)$ converge en loi vers $g(c)$.*

Démonstration. **Preuve du point 3.** Soit φ une fonction continue bornée. On a alors $\varphi o g$ qui est une fonction bornée et continue en c, et on a donc

$$\mathbb{E}\left[\varphi\left(g\left(X_n\right)\right)\right] = \mathbb{E}\left[\left(\varphi o g\right)\left(X_n\right)\right] \xrightarrow[n\to+\infty]{} \mathbb{E}\left[\left(\varphi o g\right)c\right] = \varphi(g(c)).$$

Preuve du point 2 (version 1). On a vu que pour montrer qu'une suite de variables aléatoires converge en probabilité vers une constante, il suffit de montrer qu'elle converge en loi, ce que nous donne le point 1.

Preuve du point 2 (version 2). Comme g est continue en c, pour tout $\epsilon > 0$ il existe $\eta > 0$ tel que $|x - c| \leq \eta \Longrightarrow |g(x) - g(c)| \leq \epsilon$. Ainsi,

$$\mathbb{P}\left[\left|g\left(X_n\right) - g(c)\right| > \epsilon\right] \leq \mathbb{P}\left[\left|X_n - c\right| > \eta\right] \xrightarrow[n\to+\infty]{} 0.$$

Preuve du point 1. On note

$$\Omega' = \left\{\omega \in \Omega, X_n(\omega) \xrightarrow[n\to+\infty]{} c\right\}$$

et on rappelle que par définition de la convergence presque sûre, on à $\mathbb{P}\left[\Omega'\right] = 1$. De plus, pour tout $\omega \in \Omega'$, on a par continuité

$$g\left(X_n(\omega)\right) \xrightarrow[n\to+\infty]{} g(c).$$

□

Remarque : Attention, le théorème de continuité n'est pas vrai pour la convergence en moyenne quadratique. Le problème vient en partie du fait que $\mathbb{E}\left[g\left(X_n\right)\right]$ (où sa limite) n'est pas définie.

Contre-exemple : On considère une suite de variable aléatoire (X_n) telle que $\mathbb{P}\left[X_n = 0\right] = 1 - \frac{1}{n^3}$ et $\mathbb{P}\left[X_n = n\right] = \frac{1}{n^3}$. On a

$$\mathbb{E}\left[X_n^2\right] = \frac{1}{n} \xrightarrow[n\to+\infty]{} 0$$

et X_n converge donc en moyenne quadratique vers 0. On considère maintenant la fonction $g : x \longmapsto x^2$, on a

$$\mathbb{E}\left[\left(g\left(X_n\right) - g(0)\right)^2\right] = \mathbb{E}\left[X_n^4\right] = n \xrightarrow[n\to+\infty]{} +\infty.$$

Exemple : Reprenons l'exemple de la pièce avec $\theta \in (0,1)$. On considère la fonction $g : x \longmapsto \frac{1}{\sqrt{x(1-x)}}$ qui est continue en θ. On obtient donc, par le théorème de continuité,

$$\frac{1}{\sqrt{\hat{\theta}_n\left(1-\hat{\theta}_n\right)}} \xrightarrow[n\to+\infty]{p.s} \frac{1}{\sqrt{\theta(1-\theta)}}.$$

En particulier, comme la multiplication par une constante est également une fonction continue, on obtient

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n\left(1-\hat{\theta}_n\right)}} \xrightarrow[n\to+\infty]{p.s} 1.$$

Attention ! Le théorème de continuité ne nécessite pas que g soit continue voire même définie en X_n. Dans l'exemple précédent, si on note $g : x \longleftarrow \frac{1}{x(1-x)}$, elle n'est pas définie en 0 et 1 et donc pas nécessairement définie en $\hat{\theta}_n$. Cependant, on peut remarquer

$$\mathbb{P}\left[\hat{\theta}_n = 0\right] = \mathbb{P}\left[\forall i, X_i = 0\right] = (1-\theta)^n \qquad \text{et} \qquad \mathbb{P}\left[\hat{\theta}_n = 1\right] = \mathbb{P}\left[\forall i, X_i = 1\right] = \theta^n.$$

Remarque 1.5.1. *On peut donner une version moins restrictive du théorème de continuité, mais peut-être un peu moins digeste : soit (X_n) une suite de variables aléatoires et soit X une variable aléatoire. Soit g une fonction dont l'ensemble des points de discontinuité est noté D_g. Si $\mathbb{P}\left[X \in D_g\right] = 0$, alors la suite $g(X_n)$ hérite du mode de convergence de la suite (X_n), i.e*

1. *Si (X_n) converge presque sûrement vers X, alors $g(X_n)$ converge presque sûrement vers $g(X)$.*
2. *Si (X_n) converge en probabilité vers X, alors $g(X_n)$ converge en probabilité vers $g(X)$.*
3. *Si (X_n) converge en loi vers X, alors $g(X_n)$ converge en loi vers $g(X)$.*

Afin d'établir un TCL, il ne nous reste plus qu'à utiliser le théorème de Slutsky suivant :

Théorème 1.5.2 (Théorème de Slutsky). *Soient $(X_n), (Y_n)$ des suites de variables aléatoires telles que (X_n) converge en loi vers X et (Y_n) converge en probabilité vers une constante c, alors*

$$X_n + Y_n \xrightarrow[n\to+\infty]{\mathcal{L}} X + c \qquad\qquad X_n Y_n \xrightarrow[n\to+\infty]{\mathcal{L}} cX.$$

Remarque : Le théorème précédent est généralement faux si Y_n converge en loi vers une variable aléatoire Y.
Contre-exemple : Soit $X \sim \mathcal{N}(0,1)$ et $(X_n), (Y_n)$ deux suites de variables aléatoires telles que pour tout $n \geq 1$, $X_n = -X$ et $Y_n = X$. On a

$$X_n + Y_n \xrightarrow[n\to+\infty]{\mathcal{L}} 0 \neq 2X$$

et

$$X_n Y_n \xrightarrow[n\to+\infty]{\mathcal{L}} -X^2 \neq X^2.$$

Exemple : Reprenons l'exemple du lancer de pièce et rappelons que nous avons obtenu le TLC suivant :

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

Nous avons vu que $\frac{1}{\sqrt{\theta(1-\theta)}}$ est inconnu, ce qui pourrait (par la suite) nous empêcher de construire des intervalles de confiance. Naturellement, on va plutôt s'intéresser à la variable suivante

$$\sqrt{n}\frac{1}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}\left(\hat{\theta}_n - \theta\right),$$

que l'on peut également écrire comme

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}}\sqrt{n}\frac{1}{\sqrt{\theta(1-\theta)}}\left(\hat{\theta}_n-\theta\right).$$

Comme on a

$$\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\hat{\theta}_n(1-\hat{\theta}_n)}} \xrightarrow[n\to+\infty]{p.s} 1 \quad \text{et} \quad \sqrt{n}\frac{1}{\sqrt{\theta(1-\theta)}}\left(\hat{\theta}_n-\theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1),$$

en appliquant le Théorème de Slutsky (car la convergence p.s implique la convergence en probabilité), on obtient

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

Corollaire 1.5.1. *Soit (X_n) une suite de variables aléatoires, a et σ^2 des constantes telles que*

$$\sqrt{n}\left(X_n-a\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,\sigma^2\right),$$

alors (X_n) converge en probabilité vers a.

Démonstration. On va montrer, à l'aide du théorème de Slutsky, la convergence en loi de (X_n) vers a (qui implique alors la convergence en probabilité)

$$\left(X_n-a\right)=\frac{1}{\sqrt{n}}\sqrt{n}\left(X_n-a\right),$$

et comme $\frac{1}{\sqrt{n}}$ converge en probabilités vers 0, en appliquant Slutsky, on obtient le résultat. □

Remarque 1.5.2. *Notons que le corollaire précédent peut être généralisé, i. e on peut l'écrire comme suit : soit (v_n) une suite déterministe tendant vers $+\infty$, (X_n) une suite de variables aléatoires, X une variable aléatoire et a une constante telles que*

$$v_n\left(X_n-a\right) \xrightarrow[n\to+\infty]{\mathcal{L}} X$$

alors

$$X_n \xrightarrow[n\to+\infty]{\mathbb{P}} a.$$

Théorème 1.5.3 (Delta méthode). *Soit (X_n) une suite de variables aléatoires, a, σ^2 des constantes telles que*

$$\sqrt{n}\left(X_n-a\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,\sigma^2\right),$$

alors

$$\sqrt{n}\left(g\left(X_n\right)-g\left(a\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,\left(g'(a)\right)^2\sigma^2\right).$$

Démonstration. D'après le Corollaire 1.5.1, on a

$$X_n \xrightarrow[n\to+\infty]{\mathbb{P}} a.$$

De plus, comme g est dérivable en a, il existe (développement de Taylor) une fonction r telle que

$$g(x) = g(a) + (x-a)(g'(a) + r(x))$$

avec $r(a) = 0$ et r continue en a. Plus précisément, on considère la fonction définie pour tout $x \neq a$ par

$$r(x) = \frac{g(x) - g(a) - (x-a)g'(a)}{x-a}$$

et $r(a) = 0$. En particulier, comme X_n converge en probabilité vers a, en appliquant le théorème de continuité, on obtient que $r(X_n)$ converge en probabilité vers $r(a) = 0$. De plus,

$$g(X_n) = g(a) + (X_n - a)\left(g'(a) + r(X_n)\right),$$

ce que l'on peut écrire comme

$$\sqrt{n}\left(g(X_n) - g(a)\right) = \left(g'(a) + r(X_n)\right)\sqrt{n}\left(X_n - a\right).$$

Comme

$$\left(g'(a) + r(X_n)\right) \xrightarrow[n\to+\infty]{\mathbb{P}} g'(a) \quad \text{et} \quad \sqrt{n}(X_n - a) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

on obtient le résultat en appliquant le Théorème de Slutsky. □

Exemple : Reprenons l'exemple du lancer de pièce. Supposons que l'on s'intéresse à l'estimation de $\frac{1}{\theta}$, avec $\theta > 0$. La fonction $g : x \longmapsto \frac{1}{x}$ est continue en θ et par le théorème de continuité, on a donc

$$\frac{1}{\hat{\theta}_n} \xrightarrow[n\to+\infty]{p.s} \frac{1}{\theta}.$$

De plus, comme

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

et comme g est dérivable en θ, on obtient

$$\sqrt{n}\left(\frac{1}{\hat{\theta}_n} - \frac{1}{\theta}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1-\theta}{\theta^3}\right).$$

Remarque 1.5.3. *A noter que là aussi on a donné une version simplifiée de la Delta-méthode. En effet, on peut la réécrire comme : soit (X_n) une suite de variables aléatoires, (v_n) une suite de réels tendant vers $+\infty$,*

X une variable aléatoire et a une constante telles que :

$$\sqrt{v_n}\left(X_n - a\right) \xrightarrow[n\to+\infty]{\mathcal{L}} X.$$

Soit g une fonction dérivable en a, alors

$$\sqrt{v_n}\left(g\left(X_n\right) - g\left(a\right)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} g'(a)X.$$

Chapitre 2

Estimation

Dans ce qui suit, on considère que l'on dispose de n données réelles $x_1, \ldots, x_n$ qui sont les mesures d'une variable quantitative. De plus, on supposera que ces données sont les réalisations de n variables aléatoires $X_1, \ldots, X_n$ que l'on suppose indépendantes et de même loi. On s'intéresse à une caractéristique de la loi des X_i, (espérance, variance,...) ou bien à un paramètre de cette loi, lorsque celle-ci est paramétrée (loi de Bernoulli, loi normale, loi exponentielle,...). On notera θ cette caractéristique ou ce paramètre supposé inconnu. En particulier, on s'intéressera aux deux exemples suivants :
Exemple 1 : le lancer de pièce. On considère des variables aléatoires X_i i.i.d suivant une loi de Bernoulli de paramètre θ. On s'intéressera à l'estimation du paramètre $\theta = \mathbb{E}[X_1]$.
Exemple 2 : la loi exponentielle. On considère des variables aléatoires X_i i.i.d suivant une loi exponentielle de paramètre θ. On s'intéressera à l'estimation du paramètre $\theta = (\mathbb{E}[X_1])^{-1}$.

2.1 Définitions

Dans ce qui suit, on considère un échantillon $\mathbf{X} = (X_1, \ldots, X_n)$ dépendant d'un paramètre $\theta \in \Theta$, où Θ est un ouvert de $\mathbb{R}$.

Definition 2.1.1 (Statistique et estimateur)**.** *Une statistique $T(\mathbf{X})$ est une fonction mesurable de l'échantillon $\mathbf{X}$ ne dépendant pas de θ (mais dépendant éventuellement de paramètres connus). Un estimateur de θ est une statistique $\hat{\theta} = \theta(\mathbf{X})$ destinée à approcher θ.*

Exemple : Si on prend l'exemple du lancer de pièce, la variable

$$S_n = X_1 + \ldots + X_n$$

est une statistique, mais pas un estimateur de θ contrairement à S_n/n.
Remarque : Attention ! Ne pas confondre estimateur et estimation. Dans l'exemple précédent, on a S_n qui suit une loi binomiale de paramètre n, θ. On réalise 10 lancers de pièces et on obtient 3 "Pile" et 7 "Face". On obtient donc la réalisation $s_n = 3$ et l'estimation $\theta_n = 0.3$ (et pas $S_n = 3$ et $\hat{\theta}_n = 0.3$).

On verra par la suite qu'il peut exister plusieurs estimateurs d'un même paramètre. Une manière de comparer ces estimateurs est de comparer leur risque empirique ou erreur quadratique moyenne.

Definition 2.1.2 (Erreur quadratique moyenne). *On suppose que θ est à valeurs dans $\Theta \subset \mathbb{R}$. L'erreur quadratique moyenne (ou risque quadratique) de l'estimateur $\hat{\theta}_n$ est définie pour tout $\theta \in \Theta$ par*

$$EQM\left(\hat{\theta}_n, \theta\right) = \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right]$$

Notons que si on applique l'inégalité de Markov, on obtient pour toute constante $c > 0$,

$$\mathbb{P}\left[\left|\hat{\theta}_n - \theta\right| \geq c\right] \leq \frac{\mathrm{EQM}\left(\hat{\theta}_n, \theta\right)}{c^2},$$

et donc, plus l'erreur quadratique moyenne est faible, plus la probabilité que $\hat{\theta}_n$ soit proche de θ est proche de 1. Pour comparer des estimateurs d'un même paramètre, on peut alors comparer leur erreur quadratique moyenne, et ce, pour toutes les valeurs possibles de θ.

Definition 2.1.3 (Biais d'un estimateur). *On appelle biais d'un estimateur $\hat{\theta}_n$ de θ la quantité*

$$B\left(\hat{\theta}_n, \theta\right) = \mathbb{E}\left[\hat{\theta}_n\right] - \theta.$$

1. *S'il est nul, on dit que l'estimateur est sans biais ou non biaisé.*
2. *Si $\lim_{n\to\infty} B\left(\hat{\theta}_n, \theta\right) = 0$, on dit que l'estimateur est asymptotiquement sans biais.*

Le biais permet de mesurer l'erreur moyenne de l'estimateur $\hat{\theta}_n$.

Exemple : Si on reprend l'exemple du lancer de pièce, on a vu que $\mathbb{E}\left[\hat{\theta}_n\right] = \theta$ et $\hat{\theta}_n$ est donc un estimateur sans biais de θ.

La décomposition suivante nous donne le lien entre le biais et l'erreur quadratique moyenne.

Proposition 2.1.1 (Décomposition Biais-Variance). *On a*

$$EQM\left(\hat{\theta}_n, \theta\right) = B\left(\hat{\theta}_n, \theta\right)^2 + \mathbb{V}\left[\hat{\theta}_n\right].$$

Démonstration. On a la décomposition suivante :

$$\begin{aligned}\left(\hat{\theta}_n - \theta\right)^2 &= \left(\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right) - \left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right)\right)^2 \\ &= \left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2 - 2\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)\left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right) + \left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2.\end{aligned}$$

En passant à l'espérance et par linéarité

$$\begin{aligned}\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) &= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2 - 2\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)\left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right) + \left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2\right] - 2\mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]\right)\right]\left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right) + \mathbb{E}\left[\left(\theta - \mathbb{E}\left[\hat{\theta}_n\right]\right)^2\right] \\ &= \mathbb{V}\left[\hat{\theta}_n\right] + B\left(\hat{\theta}_n, \theta\right)^2.\end{aligned}$$

□

Exemple : Prenons l'exemple du lancer de pièce, on a vu que $\hat{\theta}_n$ est un estimateur sans biais de θ. On obtient donc

$$\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) = \mathbb{V}\left[\hat{\theta}_n\right] = \frac{\theta(1-\theta)}{n}.$$

Cependant, il n'est pas toujours possible (ou facile) d'obtenir l'erreur quadratique moyenne d'un estimateur. Pour s'en convaincre, on peut considérer une variable aléatoire suivant une loi exponentielle de paramètre θ, proposer ensuite un estimateur de θ et essayer de calculer l'erreur quadratique moyenne. Il est alors important d'avoir d'autres critères pour comparer des estimateurs.

Definition 2.1.4 (Convergence, consistance et normalité asymptotique). *On dit que l'estimateur $\hat{\theta}_n$ est*

1. *convergent ou consistant si*
$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta,$$
2. *fortement consistant si*
$$\hat{\theta}_n \xrightarrow[n\to+\infty]{p.s} \theta,$$
3. *asymptotiquement normal si il existe $\sigma^2 > 0$ tel que*
$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right).$$

Remarque : A l'aide des nombreuses méthodes données précédemment (Slutsky, Delta méthode,....), on peut souvent montrer la normalité asymptotique des estimateurs. Ainsi, pour comparer deux estimateurs, on peut vérifier que ceux-ci convergent à la même vitesse avant de comparer les variances asymptotiques obtenues.

2.2 Estimation de la moyenne et de la variance

2.2.1 Estimation de la moyenne

On considère $X_1, \ldots, X_n$ des variables aléatoires indépendantes et de même loi, d'espérance $\mu = \mathbb{E}[X_1]$. Un estimateur naturel de la moyenne est donc

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

Proposition 2.2.1.

1. *$\overline{X}_n$ est un estimateur sans biais et (fortement) consistant de μ.*
2. *Si $\sigma^2 = \mathbb{V}[X_1] < +\infty$, alors $\overline{X}_n$ est asymptotiquement normal et*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right),$$
$$EQM\left(\overline{X}_n, \mu\right) = \mathbb{V}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

Dans la figure ci dessous, on voit que l'estimateur de la moyenne est bien convergent. De plus, on voit que les estimations sont centrées en la moyenne, ce qui illustre le fait que l'estimateur soit sans biais.

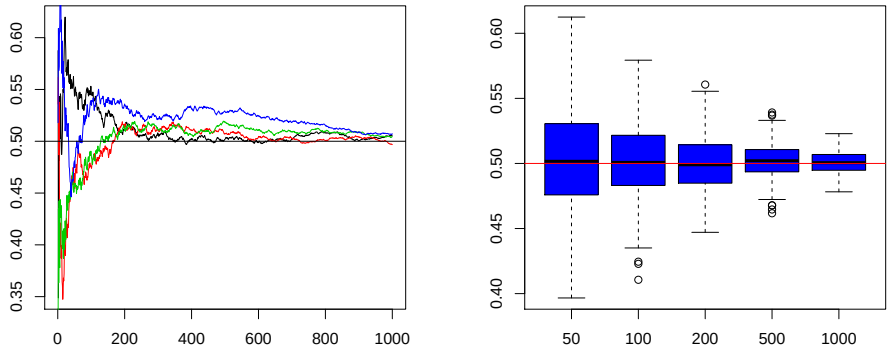

FIGURE 2.1 – Evolution de $\overline{x}_n$ par rapport à n pour 4 échantillons (à gauche) et boxplots pour les $\overline{x}_n$ obtenus pour $n = 50, 100, 200, 500, 1000$ à l'aide de 4000 échantillons (à droite).

2.2.2 Estimation de la variance

Lorsque μ est connu, un estimateur naturel de la variance est

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

et $\hat{V}_n$ est alors un estimateur sans biais et fortement consistant de σ^2 car on peut alors le voir comme l'estimateur de la moyenne de la variable $Y = (X-\mu)^2$. Cependant, μ est généralement inconnu, et on le remplace alors par $\overline{X}_n$, obtenant ainsi la variance empirique

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2}{n}\sum_{i=1}^{n}X_i\overline{X}_n + \frac{1}{n}\sum_{i=1}^{n}\overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \overline{X}_n^2.$$

Proposition 2.2.2. *$\hat{\sigma}_n^2$ est un estimateur biaisé de σ^2 mais asymptotiquement sans biais. Plus précisément,*

on a

$$\mathbb{E}\left[\hat{\sigma}_n^2\right] = \frac{n-1}{n}\sigma^2.$$

Démonstration. On a

$$\mathbb{E}\left[\hat{\sigma}_n^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[\overline{X}_n^2\right] = \mathbb{E}\left[X_1^2\right] - \mathbb{E}\left[\overline{X}_n^2\right]$$

Or,

$$\mathbb{E}\left[X_1^2\right] = \mathbb{V}\left[X_1\right] + \mathbb{E}\left[X_1\right]^2 = \sigma^2 + \mu^2,$$
$$\mathbb{E}\left[\overline{X}_n^2\right] = \mathbb{V}\left[\overline{X}_n\right] + \mathbb{E}\left[\overline{X}_n\right]^2 = \frac{1}{n}\sigma^2 + \mu^2,$$

ce qui conclut la preuve. □

Definition 2.2.1. *L'estimateur non biaisé de la variance S_n^2 est défini par*

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

Dans la figure ci-dessous, on voit que les estimations biaisées de la variance ne sont pas centrées en la variance, ce qui illustre le fait que l'estimateur soit biaisé, et ce décalage s'atténue (voire disparait) lorsque la taille d'échantillon augmente, ce qui illustre le fait que l'estimateur soit asymptotiquement sans biais. Enfin, on voit que les estimations non biaisées de la variance sont bien centrées en la variance.

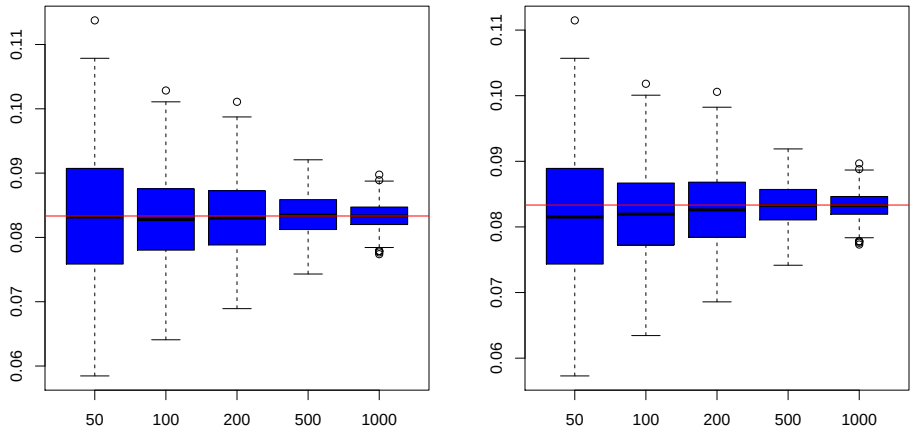

FIGURE 2.2 – Boxplots pour l'estimateur non biaisé (à gauche) et biaisé (à droite)

Proposition 2.2.3. *Les estimateurs $\hat{\sigma}_n^2$ et S_n^2 de σ^2 sont consistants.*

Démonstration. Rappelons que $\hat{\sigma}_n^2$ peut s'écrire

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

Par la loi des grands nombres, on a

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to+\infty]{\mathbb{P}} E\left[X_1^2\right] \qquad \text{et} \qquad \overline{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \mu,$$

et on obtient la consistance de $\hat{\sigma}_n^2$ à l'aide du théorème de Slutsky. De plus, on a

$$S_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2$$

et on conclut là aussi grâce au théorème de Slutsky. □

Proposition 2.2.4. *Si X_1 admet un moment d'ordre 4, alors $\hat{\sigma}_n^2$ et S_n^2 sont asymptotiquement normaux, et on a*

$$\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \tau^4 - \sigma^4\right) \quad \textit{et} \quad \sqrt{n}\left(S_n^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \tau^4 - \sigma^4\right),$$

où

$$\tau^4 = \mathbb{E}\left[(X_1 - \mu)^4\right].$$

Démonstration. Posons $Y_i = X_i - \mu$. On a alors $\overline{Y}_n = \overline{X}_n - \mu$ et

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

On peut donc écrire

$$\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \sigma^2\right) - \sqrt{n}\overline{Y}_n^2.$$

On va donc donner les convergence des deux termes à droite de l'égalité précédente. Notons d'abord que par la loi des grands nombres, $\overline{Y}_n$ converge en probabilité vers 0 (car les Y_i sont d'espérance nulle). De plus, on peut écrire

$$\sqrt{n}\overline{Y}_n^2 = \left(\sqrt{n}\overline{Y}_n\right)\overline{Y}_n.$$

Le TCL appliqué au premier terme à droite de l'égalité précédente nous donne

$$\sqrt{n}\overline{Y}_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right),$$

et comme $\overline{Y}_n$ converge en probabilité vers 0, en appliquant le Théorème de Slutsky, on obtient

$$\sqrt{n}\overline{Y}_n^2 \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

En appliquant le TCL à $\frac{1}{n}\sum_{i=1}^n Y_i^2$, on obtient

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \tau^4 - \sigma^4\right),$$

et donc, en appliquant Slutsky,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \sigma^2\right) - \sqrt{n}\overline{Y}_n^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \tau^4 - \sigma^4\right).$$

Pour l'estimateur non biaisé, comme $S_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2$, on a

$$\sqrt{n}\left(S_n^2 - \sigma^2\right) = \sqrt{n}\left(\frac{n}{n-1}\hat{\sigma}_n^2 - \sigma^2\right) = \sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) + \frac{1}{n}\sqrt{n}\hat{\sigma}_n^2$$

Pour avoir la normalité asymptotique, il faut donc montrer que le deuxième terme à droite de l'égalité précédente converge en probabilité vers 0. Notons que l'on peut écrire

$$\frac{1}{n}\sqrt{n}\hat{\sigma}_n^2 = \frac{1}{n}\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) + \frac{1}{\sqrt{n}}\sigma^2.$$

De plus, $\frac{1}{\sqrt{n}}\sigma^2$ converge vers 0, et par Slutsky,

$$\frac{1}{n}\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right) \xrightarrow[n\to+\infty]{\mathbb{P}} 0.$$

□

2.3 Méthode des moments

La méthode des moments repose sur la proposition suivante.

Proposition 2.3.1. *Soit Θ un intervalle ouvert de $\mathbb{R}$ et $\theta \in \Theta$. Soit φ un C^1-difféomorphisme de Θ dans $\varphi(\Theta)$, i.e une bijection de Θ dans $\varphi(\Theta)$, continûment dérivable et de réciproque continûment dérivable. Soit $\hat{\varphi}_n$ un estimateur consistant de $\varphi(\theta)$, alors $\hat{\theta}_n = \varphi^{-1}(\hat{\varphi}_n)$ est un estimateur consistant de θ, i.e*

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta.$$

De plus, si $\hat{\varphi}_n$ est un estimateur asymptotiquement normal de $\varphi(\theta)$, i.e si il existe $\sigma^2 > 0$ tel que

$$\sqrt{n}\left(\hat{\varphi}_n - \varphi(\theta)\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right),$$

et si $\varphi'(\theta) \neq 0$, alors $\hat{\theta}_n$ est un estimateur asymptotiquement normal de θ, et plus précisément

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{\left(\varphi'(\theta)\right)^2}\right).$$

Remarque : La propriété précédente reste vraie si on remplace $\sqrt{n}$ par une suite v_n tendant vers $+\infty$.

Démonstration. Remarquons que comme $\hat{\varphi}_n$ converge en probabilité vers $\varphi(\theta)$ et comme Θ est un ouvert

$$\mathbb{P}\left[\hat{\varphi}_n \notin \varphi\left(\Theta\right)\right] \xrightarrow[n\to+\infty]{} 0,$$

et donc l'estimateur $\hat{\theta}_n = \varphi^{-1}\left(\hat{\varphi}_n\right)$ est bien défini avec une probabilité tendant vers 1. De plus comme φ^{-1} est continue en $\varphi(\theta)$, le théorème de continuité nous donne la consistance de $\hat{\theta}_n$.
Supposons maintenant que la normalité asymptotique est vérifiée, on va appliquer la delta méthode avec $g = \varphi^{-1}$. Rappelons que pour tout $x \in \varphi(\Theta)$,

$$\left(\varphi^{-1}\right)'(x) = \frac{1}{\left(\varphi' o \varphi^{-1}\right)(x)},$$

et on obtient donc en particulier

$$\left(\varphi^{-1}\right)'(\varphi(\theta)) = \frac{1}{\varphi'\left(\varphi^{-1}\left(\varphi\left(\theta\right)\right)\right)} = \frac{1}{\varphi'(\theta)}.$$

En appliquant la delta méthode, on obtient donc

$$\sqrt{n}\left(\varphi^{-1}\left(\hat{\varphi}_n\right) - \varphi^{-1}\left(\varphi(\theta)\right)\right) \xrightarrow[n\to+\infty]{\mathcal{N}} \mathcal{N}\left(0, \left(\left(\varphi^{-1}\right)'\left(\varphi(\theta)\right)\right)^2 \sigma^2\right),$$

et on obtient donc

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\left(\varphi'(\theta)\right)^2}\sigma^2\right).$$

□

La méthode des moments consiste donc à trouver une fonction φ qui soit un difféomorphisme au voisinage de θ et un moment k tel que $\mathbb{E}\left[X_1^k\right] = \varphi(\theta)$. En effet, on a alors un estimateur "naturel" de $m_k = \mathbb{E}\left[X_1^k\right]$ donné par

$$\hat{m}_{n,k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

et on obtient alors l'estimateur $\hat{\theta}_n = \varphi^{-1}(\hat{m}_{n,k})$ de θ.

Exemple : la loi uniforme. Soit $X_1, ..., X_n$ des variables aléatoires i.i.d suivant une loi uniforme sur $[0, \theta^2]$, avec $\theta > 0$, et donc de densité

$$f_\theta = \frac{1}{\theta^2} \mathbb{1}_{[0,\theta^2]}.$$

On a $\mathbb{E}[X_1] = \frac{\theta^2}{2}$. Par la loi des grands nombres, $\overline{X}_n$ converge en probabilité vers $\theta^2/2$. On pose alors $\varphi : x \longmapsto x^2/2$ qui est bien un C^1-difféomorphisme de $\mathbb{R}_+^*$ dans $\mathbb{R}_+^*$, et d'inverse $\varphi^{-1} : x \longmapsto \sqrt{2x}$. On obtient alors la convergence en probabilité de $\hat{\theta}_n = \sqrt{2\overline{X}_n}$ vers θ. De plus, comme $\mathbb{V}[X_1] = \frac{\theta^4}{12}$ grâce au TLC, on a

$$\sqrt{n}\left(\overline{X}_n - \frac{\theta^2}{2}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta^4}{12}\right).$$

Comme $\varphi'(\theta) = \theta$, on obtient la normalité asymptotique suivante

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta^2}{12}\right)$$

Exemple : la loi exponentielle. On considère $X_1, \ldots, X_n$ des variables aléatoires i.i.d suivant une loi exponentielle de paramètre $\theta \in \mathbb{R}_+^*$, i.e ayant une densité définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \theta \exp(-\theta x) \mathbb{1}_{\mathbb{R}_+}(x).$$

Rappelons que $\mathbb{E}[X_1] = \theta^{-1}$ et $\mathbb{V}[X_1] = \theta^{-2}$. La loi des grands nombres nous donne la convergence en probabilité de $\overline{X}_n$ vers θ^{-1}. On pose alors $\varphi : x \longmapsto x^{-1}$ qui est un C^1 difféomorphisme de $\mathbb{R}_+^*$ dans $\mathbb{R}_+^*$. On a donc la convergence en probabilité de $\hat{\theta}_n = \overline{X}_n^{-1}$ vers θ. De plus, grâce au TCL, on a

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

Comme $\varphi'(\theta) = \frac{-1}{\theta^2}$, on obtient

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

Remarque : Pour éviter les erreurs, il est plus judicieux (et cela revient à peu près au même) d'écrire θ comme une fonction de $\mathbb{E}\left[X^k\right]$ et d'appliquer ensuite le théorème de continuité pour avoir la consistance, et la delta méthode pour obtenir la normalité asymptotique...

Remarque : Attention ! Il peut arriver qu'une variable aléatoire n'admette pas de moment d'ordre 1. Il faut alors essayer d'être malin !

Exemple : Soit $\theta > 0$, on considère une variable aléatoire X de densité définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{\theta}{x^2} \mathbf{1}_{x \geq \theta}.$$

On a ainsi

$$\mathbb{E}[X] = \int_\theta^{+\infty} \frac{\theta}{x} dx = +\infty.$$

Cependant, on peut s'intéresser à

$$\mathbb{E}\left[\frac{1}{X}\right] = \int_\theta^{+\infty} \frac{\theta}{x^3} dx = \frac{1}{2\theta}.$$

Ainsi, comme $\theta = \frac{1}{2\mathbb{E}\left[\frac{1}{X}\right]}$, on va considérer l'estimateur $\hat{\theta}_n = \frac{n}{2\sum_{i=1}^n \frac{1}{X_i}}$. Par la loi des grands nombres, on obtient

$$\frac{1}{n}\sum_{i=1}^n \frac{1}{X_i} \xrightarrow[n\to+\infty]{\mathbb{P}} \frac{1}{2\theta}$$

De plus, la fonction $\varphi : x \longmapsto \frac{1}{2x}$ est continue en $\frac{1}{2\theta} \neq 0$, et on obtient donc la consistance via le théorème de continuité. De plus,

$$\mathbb{E}\left[\frac{1}{X^2}\right] = \int_\theta^{+\infty} \frac{\theta}{x^4} dx = \frac{1}{3\theta^2}$$

et

$$\mathbb{V}\left[\frac{1}{X}\right] = \frac{1}{12\theta^2}.$$

Le TLC nous donne donc

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{X_i} - \frac{1}{2\theta}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{12\theta^2}\right).$$

De plus, $\varphi'(x) = -\frac{1}{2x^2}$, et φ est donc dérivable en $\frac{1}{2\theta}$ avec $\varphi'\left(\frac{1}{2\theta}\right) = 2\theta^2$. On obtient donc, par la delta méthode,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{\theta^2}{3}\right).$$

Remarque : De plus, même si les moments d'ordre $1, 2, ...$ existent, il n'est pas toujours facile de les calculer. Il peut alors être intéressant de calculer la fonction génératrice des moments.

Definition 2.3.1 (Fonction génératrice des moments)**.** *Soit X une variable aléatoire, on appelle fonction génératrice des moments G_X de X la fonction définie par*

$$G_X(t) = \mathbb{E}\left[e^{tX}\right].$$

Attention, la fonction génératrice des moments n'est pas nécessairement définie, et encore moins pour tout t, mais le théorème suivant nous indique quand elle peut l'être.

Théorème 2.3.1. *On suppose que la variable aléatoire X admet des moments de tout ordre, i.e pour tout $k \in \mathbb{N}$ $\mathbb{E}\left[X^k\right] < +\infty$, et que la série $\sum_{k\geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!}$ admet un rayon de convergence R. Alors pour tout*

$|t| < R$ on a

$$G_X(t) = \sum_{k\geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!}.$$

En particulier, pour tout $k \in \mathbb{N}$, on a

$$G_X^{(k)}(0) = \mathbb{E}\left[X^k\right].$$

Démonstration. On a, grâce à la décomposition en série entière de la fonction exponentielle,

$$e^{tX} = \sum_{k\geq 0} \frac{t^k X^k}{k!}$$

et on a donc, pour tout t tel que $|t| < R$, à l'aide du théorème de Fubini-Tonelli,

$$\mathbb{E}\left[e^{tX}\right] = \mathbb{E}\left[\sum_{k\geq 0} \frac{t^k X^k}{k!}\right] = \sum_{k\geq 0} \frac{t^k \mathbb{E}\left[X^k\right]}{k!}.$$

En particulier, pour tout $k \geq 0$, on a

$$G_X^{(k)}(t) = \mathbb{E}\left[X^k\right] + \sum_{k'\geq k+1} \frac{k'(k'-1)...(k'-k+1)}{k'!} t^{k'-k} \mathbb{E}\left[X^{k'}\right]$$

et en particulier, on obtient $G_X^{(k)}(0) = \mathbb{E}\left[X^k\right]$. □

Exemple : la loi géométrique. Soit X une variable aléatoire suivant une loi géométrique de paramètre $p \in (0,1)$. On a alors pour tout $t \geq 0$, en faisant le changement de variable $k' = k-1$,

$$G_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{k\geq 1} e^{tk}(1-p)^{k-1}p = e^t p \sum_{k'=0}^{+\infty} e^{tk'}(1-p)^{k'} = e^t p \sum_{k=0}^{+\infty} \left((1-p)e^t\right)^{k'}.$$

En prenant t tel que $(1-p)e^t < 1$, i.e $t < -\log(1-p)$ on a donc

$$G_X(t) = \frac{e^t p}{1-(1-p)e^t} = \frac{p}{e^{-t}-(1-p)}.$$

Ainsi, on a pour tout $t < -\log(1-p)$

$$G_X'(t) = p\frac{e^{-t}}{\left(e^{-t}-(1-p)\right)^2}$$

et on a ainsi $\mathbb{E}[X] = G_X'(0) = \frac{p}{p^2} = \frac{1}{p}$. De plus, on a

$$G_X''(t) = p\frac{-e^{-t}\left(e^{-t}-(1-p)\right)^2 + 2e^{-2t}\left(e^{-t}-(1-p)\right)}{\left(e^{-t}-(1-p)\right)^4} = p\frac{-e^{-t}\left(e^{-t}-(1-p)\right) + 2e^{-2t}}{\left(e^{-t}-(1-p)\right)^3}$$

et on a ainsi $\mathbb{E}\left[X^2\right] = G_X''(0) = p\frac{2-p}{p^3} = \frac{2-p}{p^2}$. Ainsi, on a

$$\mathbb{V}[X] = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

2.4 Méthode du maximum de vraisemblance

Dans ce qui suit, on considère des variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$.

2.4.1 Cas discret

On note f_θ la densité de X_1, i.e la fonction définie pour tout $x \in \mathbb{R}$ par $f_\theta(x) = \mathbb{P}[X_1 = x]$.

Definition 2.4.1 ((log)-Vraisemblance)**.** *La vraisemblance de* $\mathbf{X} = (X_1, \ldots, X_n)$ *est définie pour tout* $\theta \in \Theta$ *par*

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} f_\theta(X_i).$$

La log-vraisemblance de $\mathbf{X}$ *est définie pour tout* $\theta \in \Theta$ *par*

$$l_{\mathbf{X}}(\theta) = \log(L_{\mathbf{X}}(\theta) = \log\left(\prod_{i=1}^{n} f_\theta(X_i)\right).$$

Attention ! La (log)-vraisemblance est une variable aléatoire.

Exemple 1 : loi de Bernoulli. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et suivent une loi de Bernoulli de paramètre $\theta \in (0,1)$, i.e la densité de X_1 est définie pour tout $x \in \{0,1\}$ par

$$f_\theta(x) = \theta^x(1-\theta)^{1-x}.$$

On obtient donc pour tout $\theta \in (0,1)$,

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{\sum_{i=1}^n X_i}(1-\theta)^{\sum_{i=1}^n (1-X_i)} = \theta^{n\overline{X}_n}(1-\theta)^{n-n\overline{X}_n}.$$

De plus, on a

$$l_{\mathbf{X}}(\theta) = n\overline{X}_n \ln(\theta) + \left(n - n\overline{X}_n\right)\log(1-\theta).$$

Exemple 2 : loi de Poisson. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et suivent une loi de Poisson de paramètre $\theta \in \mathbb{R}_+^*$, i.e la densité de X_1 est définie pour tout $x \in \mathbb{N}$ par

$$f_\theta(x) = \frac{\theta^x}{x!}e^{-\theta}.$$

On obtient donc pour tout $\theta \in \Theta$

$$L_n(\theta) = \prod_{i=1}^{n} \frac{\theta^{X_i}}{X_i!} e^{-\theta} = e^{-\theta n} \prod_{i=1}^{n} \frac{\theta^{X_i}}{X_i!}$$
$$l_n(\theta) = \log\left(e^{-\theta n}\right) \sum_{i=1}^{n} \log\left(\frac{\theta^{X_i}}{X_i!}\right) = -\theta n + \sum_{i=1}^{n} X_i \log(\theta) - \sum_{i=1}^{n} \log\left(X_i!\right)$$

Si on note $x_1, \ldots, x_n$ les réalisations de $X_1, \ldots, X_n$, la réalisation de la vraisemblance est la probabilité d'obtenir une telle réalisation de l'échantillon si θ est le vrai paramètre qui régit X_1. L'objectif est de trouver θ qui va maximiser cette probabilité, ce qui conduit à l'estimateur du maximum de vraisemblance.

Definition 2.4.2 (Estimateur du maximum de vraisemblance (EMV))**.** *Le maximum de vraisemblance, si il existe, est un élément $\hat{\theta}_n$ de Θ qui vérifie*

$$L_{\mathbf{X}}\left(\hat{\theta}_n\right) = \sup_{\theta \in \Theta} L_{\mathbf{X}}(\theta).$$

De manière équivalente, l'estimateur du maximum de vraisemblance, si il existe, vérifie

$$l_{\mathbf{X}}\left(\hat{\theta}_n\right) = \sup_{\theta \in \Theta} l_{\mathbf{X}}(\theta).$$

On voit donc que la réalisation θ_n de l'estimateur $\hat{\theta}_n$ maximise la probabilité d'obtenir un tel échantillon.

Attention ! Ni l'existence ni l'unicité de l'estimateur du maximum de vraisemblance ne sont assuré.

Exemple : loi de Bernoulli. On va chercher à minimiser la log-vraisemblance. En notant $l'_{\mathbf{X}}(\theta) = \frac{\partial}{\partial\theta} l_{\mathbf{X}}(\theta)$, on a

$$l'_{\mathbf{X}}(\theta) = \frac{n\overline{X}_n}{\theta} - \frac{n - n\overline{X}_n}{1-\theta} = \frac{n\theta - n\overline{X}_n}{\theta(1-\theta)} = 0 \Leftrightarrow \theta = \overline{X}_n.$$

De plus, en dressant le tableau de variation, on voit que $\overline{X}_n$ est l'unique maximiseur de la log-vraisemblance, et on a donc l'EMV $\hat{\theta}_n^{MV} = \overline{X}_n$.

2.4.2 Cas continu

On suppose maintenant que la variable aléatoire X_1 est continue et admet une densité f_θ (par rapport à la mesure de Lebesgue). La définition de la (log)-vraisemblance reste inchangée.

Definition 2.4.3 ((log)-Vraisemblance)**.** *La vraisemblance de $\mathbf{X} = (X_1, \ldots, X_n)$ est définie pour tout $\theta \in \Theta$ par*

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} f_\theta\left(X_i\right).$$

La log-vraisemblance de $\mathbf{X}$ *est définie pour tout* $\theta \in \Theta$ *par*

$$l_{\mathbf{X}}(\theta) = \log\left(L_{\mathbf{X}}(\theta) = \log\left(\prod_{i=1}^{n} f_\theta\left(X_i\right)\right)\right.\right.$$

La définition de l'estimateur du maximum de vraisemblance reste donc inchangée dans le cas continue. Cependant, on peut noter que dans ce cas, la réalisation de la vraisemblance est la densité en la réalisation, et la réalisation de l'estimateur du maximum de vraisemblance est donc le paramètre θ qui maximise la densité en l'échantillon.

Exemple 1 : loi normale. On considère $\mathbf{X} = (X_1, \ldots, X_n)$ où les X_i sont i.i.d et suivent une loi normale d'espérance $\theta \in \mathbb{R}$ et de variance 1, i.e la densité de X_1 est définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right).$$

On obtient donc pour tout $\theta \in \Theta$

$$L'_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i-\theta)^2}{2}\right) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2\right)$$
$$l'_{\mathbf{X}}(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2$$

En dérivant par rapport à θ, on obtient

$$l'_{\mathbf{X}}(\theta) = \sum_{i=1}^{n} \frac{X_i - \theta}{\sigma^2}.$$

On cherche alors les zéros de la dérivée, et en multipliant par σ^2, on obtient

$$\begin{aligned}\sum_{i=1}^{n} \frac{X_i-\theta}{\sigma^2} = 0 &\Leftrightarrow \sum_{i=1}^{n}(X_i-\theta) = 0\\ &\Leftrightarrow n\overline{X}_n - n\overline{\theta} = 0\\ &\Leftrightarrow \theta = \overline{X}_n.\end{aligned}$$

Reste à montrer que $\hat{\theta}_n = \overline{X}_n$ est bien l'unique maximum de la log vraisemblance. On regarde donc la dérivée seconde par rapport à θ, et on obtient

$$l''_{\mathbf{X}}(\theta) = \frac{-n}{\sigma^2} < 0$$

et la log-vraisemblance est donc strictement concave, et $\hat{\theta}_n$ est son unique maximum. On obtient donc l'EMV

$$\hat{\theta}_n = \overline{X}_n.$$

Exemple 2 : On considère que X_1 admet une densité f_θ définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{\theta}{x^2}\mathbf{1}_{x\geq\theta}.$$

avec $\theta > 0$. On a alors la vraisemblance qui est définie pour tout $\theta > 0$ par

$$L_n(\theta) = \prod_{i=1}^n \frac{\theta^n}{X_i^2}\mathbf{1}_{X_i\geq\theta} = \theta^n \prod_{i=1}^n \frac{1}{X_i^2}\prod_{i=1}^n \mathbf{1}_{X_i\geq\theta}.$$

De plus

$$\prod_{i=1}^n \mathbf{1}_{X_i\geq\theta} = 1 \Leftrightarrow \forall i, X_i \geq \theta \Leftrightarrow \theta \leq X_{(1)} := \min_i X_i.$$

On a alors

$$L_n(\theta) = \theta^n \mathbf{1}_{\theta\leq X_{(1)}} \prod_{i=1}^n \frac{1}{X_i^2}$$

qui est nulle sur $\left(X_{(1)}, +\infty\right)$, et positive et croissante sur $\left(0, X_{(1)}\right]$, et le maximum est donc atteint en $X_{(1)}$ et est unique. On a donc l'estimateur du maximum de vraisemblance $\hat{\theta}_n = X_{(1)}$. Il reste alors à démontrer la convergence de l'estimateur. Comme $\hat{\theta}_n$ ne peut pas être vu comme une fonction de $\overline{X}_n$, on ne peut pas utiliser le cheminement classique, LGN, théorème de continuité, TLC, delta méthode... Cependant, pour tout $\epsilon > 0$, comme $X_{(1)} \geq \theta$,

$$\mathbb{P}\left[\left|X_{(1)} - \theta\right| \geq \epsilon\right] = \mathbb{P}\left[X_{(1)} - \theta \geq \epsilon\right] = \mathbb{P}\left[X_{(1)} \geq \epsilon + \theta\right] = \mathbb{P}\left[\forall i, X_i \geq \theta + \epsilon\right].$$

De plus, comme les X_i sont indépendants et identiquement distribués, on obtient

$$\mathbb{P}\left[\left|X_{(1)} - \theta\right| \geq \epsilon\right] = \prod_{i=1}^n \mathbb{P}\left[X_i \geq \theta + \epsilon\right] = \left(\mathbb{P}\left[X_1 \geq \theta + \epsilon\right]\right)^n = \left(1 - F_{X_1}(\theta + \epsilon)\right)^n,$$

où F_{X_1} est la fonction de répartition de X_1 qu'il faut donc calculer. On a pour tout $x \geq \theta$

$$F_{X_1}(x) = \int_\theta^x \frac{\theta}{t^2} dt = \theta\left[\frac{-1}{t}\right]_\theta^x = 1 - \frac{\theta}{x}.$$

On obtient donc

$$\mathbb{P}\left[\left|X_{(1)} - \theta\right| \geq \epsilon\right] = \left(\frac{\theta}{\theta + \epsilon}\right)^n \xrightarrow[n\to+\infty]{} 0.$$

Reste encore à trouver la vitesse de convergence. Pour "plagier" le TLC, on va considérer une suite positive $v_n \to +\infty$ telle que

$$V_n := v_n\left(X_{(1)} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} Z$$

où Z suit une loi connue (si possible). Pour obtenir la convergence en loi, on va donc regarder la fonction de répartition de V_n. Comme $X_{(1)} \geq \theta$, pour tout $t < 0$, on a $\mathbb{P}\left[V_n \leq t\right] = 0$. Pour tout

$t \geq 0$, on a (en utilisant les calculs précédents)

$$\mathbb{P}\left[V_n \leq t\right] = \mathbb{P}\left[X_{(1)} - \theta \leq \frac{t}{v_n}\right] = 1 - \mathbb{P}\left[X_{(1)} - \theta \geq \frac{t}{v_n}\right] = 1 - \left(\frac{\theta}{\theta + t/v_n}\right)^n.$$

On obtient donc

$$\mathbb{P}\left[V_n \leq t\right] = 1 - \left(1 - \frac{t/v_n}{\theta + t/v_n}\right)^n = 1 - \left(1 - \frac{t}{\theta v_n + t}\right)^n = 1 - \exp\left(n \ln\left(1 - \frac{t}{\theta v_n + t}\right)\right) \sim 1 - \exp\left(-\frac{tn}{\theta v_n + t}\right)$$

En choisissant $v_n = n$, on a

$$\mathbb{P}\left[V_n \leq t\right] \xrightarrow[n\to+\infty]{} 1 - \exp\left(-\frac{t}{\theta}\right),$$

i.e on a la convergence en loi

$$n\left(X_{(1)} - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{E}\left(\frac{1}{\theta}\right).$$

2.5 Comparaison d'estimateurs

On a vu, dans les sections précédentes, différentes façons de construire des estimateurs. On s'intéresse ici à la façon de les comparer et donc comment choisir le meilleur (si il existe).

2.5.1 Comparaison des erreurs quadratiques moyennes

Une façon de quantifier la qualité d'un estimateur $\hat{\theta}_n$ de θ, est de considérer son risque quadratique

$$\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) = \mathbb{E}\left[\left(\hat{\theta}_n - \theta\right)^2\right].$$

On considère que $\hat{\theta}_n$ est un meilleur estimateur de θ que l'estimateur $\tilde{\theta}_n$ si

$$\forall \theta \in \Theta, \qquad \mathrm{EQM}\left(\hat{\theta}_n, \theta\right) \leq \mathrm{EQM}\left(\tilde{\theta}_n, \theta\right).$$

Exemple : loi uniforme. Soit $\theta > 0$, on considère une variable aléatoire X suivant une loi uniforme sur $[0, \theta]$ et on se donne $X_1, \ldots, X_n$ de même loi que X. On a $\mathbb{E}[X] = \frac{\theta}{2}$ et donc, par la méthode des moments, on obtient l'estimateur

$$\hat{\theta}_n = 2\overline{X}_n.$$

qui est un estimateur sans biais de θ. De plus, comme $\mathbb{V}\left[X_1\right] = \frac{\theta^2}{12}$, on a, grâce à la décomposition biais variance,

$$\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) = \mathbb{V}\left[\hat{\theta}_n\right] = 4\frac{\mathbb{V}\left[X_1\right]}{n} = \frac{\theta^2}{3n}.$$

On s'intéresse maintenant à l'estimateur du maximum de vraisemblance. On a la vraisemblance

qui est définie pour tout $\theta > 0$ par

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(X_i) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbf{1}_{\theta \geq X_i}.$$

Or, $\prod_{i=1}^{n} \mathbf{1}_{\theta \geq X_i} = 1 \Leftrightarrow \forall i, X_i \leq \theta \Leftrightarrow X_{(n)} \leq \theta$. On a donc

$$L_{\mathbf{X}}(\theta) = \frac{1}{\theta^n} \mathbf{1}_{\theta \geq X_{(n)}}$$

qui est nulle sur $\left(0, X_{(n)}\right)$ et positive et décroissante sur $\left[X_{(n)}, +\infty\right[$, et son maximum est daonc atteint en $X_{(n)}$ qui est donc l'estimateur du maximum de vraisemblance. De plus, pour tout $t \in [0, \theta]$,

$$\begin{aligned} F_{X_{(n)}}(t) &= \mathbb{P}\left[X_{(n)} \leq t\right] \\ &= \mathbb{P}\left[\{\forall i, X_i \leq t\}\right] \\ &= \prod_{i=1}^{n} \mathbb{P}\left[X_i \leq t\right] \\ &= \left(\mathbb{P}\left[X_1 \leq t\right]\right)^n \\ &= \left(\frac{t}{\theta}\right)^n \end{aligned}$$

On obtient donc

$$f_{X_{(n)}}(t) = \frac{n}{\theta} \left(\frac{t}{\theta}\right)^{n-1} \mathbf{1}_{[0,\theta]}(t).$$

Ainsi,

$$\mathbb{E}\left[\hat{\theta}_n\right] = \int_0^\theta t \frac{n}{\theta} \left(\frac{t}{\theta}\right)^{n-1} dt = \frac{n}{n+1}\theta \int_0^\theta \frac{n+1}{\theta} \left(\frac{t}{\theta}\right)^n dt = \frac{n}{n+1}\theta$$

et l'estimateur est donc biaisé, avec en particulier

$$B\left(\hat{\theta}_n, \theta\right) = -\frac{1}{n+1}\theta$$

De plus, on a

$$\mathbb{E}\left[\hat{\theta}_n^2\right] = \int_0^\theta t^2 \frac{n}{\theta} \left(\frac{t}{\theta}\right)^{n-1} dt = \frac{n}{n+2}\theta^2 \int_0^\theta \frac{n+2}{\theta} \left(\frac{t}{\theta}\right)^{n+1} dt = \frac{n}{n+2}\theta^2$$

On obtient donc

$$\begin{aligned}\mathbb{V}\left[\hat{\theta}_n\right] &= \mathbb{E}\left[\hat{\theta}_n^2\right] - \left(\mathbb{E}\left[\hat{\theta}_n\right]\right)^2 \\ &= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 \\ &= \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2}\theta^2 \\ &= \frac{n}{(n+2)(n+1)^2}\theta^2.\end{aligned}$$

On a alors,

$$\mathrm{EQM}\left(\hat{\theta}_n, \theta\right) = \mathbb{V}\left[\hat{\theta}_n\right] + B\left(\hat{\theta}_n, \theta\right)^2 = \frac{n}{(n+2)(n+1)^2}\theta^2 + \frac{n^2}{(n+1)^2}\theta^2 = \frac{2\theta^2}{(n+1)(n+2)}$$

Et c'est donc un meilleur estimateur que l'estimateur des moments dès que $\frac{2}{(n+1)(n+2)} \leq \frac{1}{3n}$, i.e dès que $n \geq 2$.

2.5.2 Biais d'un estimateur

On donne souvent trop d'importance à l'absence de biais d'un estimateur, voire même on peut penser qu'un estimateur sera meilleur si il est non biaisé, ce qui est faux ! Débiaiser un estimateur ne donne pas forcément de meilleurs résultats !

Exemple : loi uniforme. Soit $\theta > 0$, on considère une variable aléatoire X suivant une loi uniforme sur $[0, \theta]$. On rappelle que l'estimateur du maximum de vraisemblance est $X_{(n)}$ et qu'il est biaisé avec

$$B\left(\hat{\theta}_n, \theta\right) = -\frac{1}{n+1}\theta.$$

On considère maintenant l'estimateur non biaisé $\tilde{\theta}_n$ défini par

$$\tilde{\theta}_n = \frac{n+1}{n}X_{(n)}.$$

On a

$$\mathbb{V}\left[\tilde{\theta}_n\right] = \frac{(n+1)^2}{n^2}\mathbb{V}\left[X_{(n)}\right] = \frac{\theta^2}{n(n+2)} = \mathrm{EQM}\left(\tilde{\theta}_n, \theta\right),$$

et on obtient donc un meilleur estimateur que $X_{(n)}$. Cependant, si on considère un estimateur de la forme $\alpha X_{(n)}$ avec α un réel positif, on obtient

$$\mathbb{E}\left[\alpha X_{(n)}\right] = \alpha\frac{n}{n+1}\theta \qquad \text{et} \qquad B\left(\alpha X_{(n)}, \theta\right) = \frac{\alpha n - n - 1}{n+1}\theta$$

et

$$\begin{aligned}\mathrm{EQM}\left(\alpha X_n, \theta\right) &= \alpha^2 \mathbb{E}\left[X_{(n)}^2\right] - 2\alpha\theta \mathbb{E}\left[X_{(n)}\right] + \theta^2 \\ &= \theta^2\left(\frac{n}{n+2}\alpha^2 - \frac{2n}{n+1}\alpha + 1\right)\end{aligned}$$

On prend alors α qui minimise la quantité précédente, i.e $\alpha = \frac{n+2}{n+1}$ et on obtient ainsi un meilleur estimateur.

2.5.3 L'approche asymptotique

On a vu qu'il n'était pas toujours évident (voire possible) de calculer les erreurs quadratiques moyennes. De plus, il est relativement atypique d'obtenir des vitesses de convergence en moyenne quadratique de l'ordre de $1/n^2$. Très souvent, on dispose d'estimateurs $\hat{\theta}_n, \tilde{\theta}_n$ asymptotiquement normaux, i.e il existe σ_1^2, σ_2^2 telles que

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma_1^2\right) \quad \text{et} \quad \sqrt{n}\left(\tilde{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma_2^2\right)$$

Si on a $\sigma_1^2 \leq \sigma_2^2$, on choisira alors l'estimateur $\hat{\theta}_n$. En effet, on verra (chapitre suivant) que l'on a alors, par exemple,

$$\mathbb{P}\left[\theta \in \left[\hat{\theta} \pm 1.96\frac{\sigma_1}{\sqrt{n}}\right]\right] \xrightarrow[n\to+\infty]{} 0.95 \quad \text{et} \quad \mathbb{P}\left[\theta \in \left[\tilde{\theta} \pm 1.96\frac{\sigma_2}{\sqrt{n}}\right]\right] \xrightarrow[n\to+\infty]{} 0.95$$

et le premier intervalle est alors plus précis. En effet, les deux intervalles sont alors de longueurs

$$2 \times 1.96\frac{\sigma_1}{\sqrt{n}} \quad \text{et} \quad 2 \times 1.96\frac{\sigma_2}{\sqrt{n}}.$$

Chapitre 3

Intervalles de confiance

Dans ce chapitre, on considère des variables aléatoires i.i.d $X_1, \ldots, X_n$ suivant une loi F. On s'intéresse à l'estimation d'une caractéristique ou d'un paramètre θ de la loi F. On a vu dans les chapitres précédents comment obtenir et comparer des estimateurs $\hat{\theta}_n$ de θ. On s'intéresse ici à la quantification de la confiance que l'on peut accorder à l'estimation, i.e à l'obtention d'intervalles de confiance.

3.1 Intervalle de confiance

3.1.1 Intervalle de confiance

Soit $\alpha \in (0,1)$, un intervalle de confiance pour le paramètre θ au niveau de confiance $1-\alpha$ est un intervalle de la forme

$$IC_{1-\alpha}(\theta) = [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$$

avec

$$\mathbb{P}[\theta \in [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]] = 1-\alpha.$$

On parle alors d'intervalles de confiance bilatères.

Attention ! Cela ne signifie pas que θ appartienne à $IC_{1-\alpha}(\theta)$. On peut seulement affirmer que $\theta \in IC_{1-\alpha}(\theta)$ en prenant un risque α.

Attention ! On ne peut en aucun cas dire que la probabilité que θ appartienne à la réalisation de $IC_{1-\alpha}(\theta)$ est égale à $1-\alpha$. En effet, la réalisation de cet intervalle n'est en aucun cas aléatoire, et donc la probabilité que θ appartienne à cet intervalle est soit 0 soit 1.

Exemple 1 : On considère des variables aléatoires $X_1, \ldots, X_n$ i.i.d de densité f_θ définie pour tout $x \in \mathbb{R}$ par

$$f_\theta(x) = \frac{\theta}{x^2}\mathbf{1}_{x \geq \theta},$$

avec $\theta \in \mathbb{R}$. On a vu que l'estimateur du maximum de vraisemblance est défini par $X_{(1)}$. Comme

$\theta \leq X_{(1)}$, on va chercher un intervalle de la forme

$$IC_{1-\alpha}(\theta) = \left[X_{(1)} - c_\alpha, X_{(1)}\right].$$

Rappelons que l'on a vu que pour tout $t \geq 0$,

$$\mathbb{P}\left[X_{(1)} - \theta \geq t\right] = \left(\frac{\theta}{\theta + t}\right)^n.$$

De plus, on cherche c_α tel que

$$\begin{aligned} 1 - \alpha = \mathbb{P}\left[X_{(1)} - c_\alpha \leq \theta \leq X_{(1)}\right] = \mathbb{P}\left[X_{(1)} - c_\alpha \leq \theta\right] &\Leftrightarrow \alpha = \mathbb{P}\left[X_{(1)} - \theta \geq c_\alpha\right] \\ &\Leftrightarrow \alpha = \left(\frac{\theta}{\theta + c_\alpha}\right)^n \\ &\Leftrightarrow \alpha^{1/n} = \frac{\theta}{\theta + c_\alpha} \\ &\Leftrightarrow c_\alpha = \frac{\theta}{\alpha^{1/n}} - \theta = \theta\left(\alpha^{-1/n} - 1\right). \end{aligned}$$

Si on se contente de cela, on a un intervalle de la forme

$$IC_{1-\alpha}(\theta) = \left[X_{(1)} - \theta\left(\alpha^{-1/n} - 1\right); X_{(1)}\right],$$

donc dépendant de θ, et on ne peut donc pas le calculer. Il faut donc continuer de travailler. On a

$$\alpha = \mathbb{P}\left[X_{(1)} - \theta \geq \theta\left(\alpha^{-1/n} - 1\right)\right] = \mathbb{P}\left[X_{(1)} \geq \theta\alpha^{-1/n}\right] = \mathbb{P}\left[X_{(1)}\alpha^{1/n} \geq \theta\right]$$

et on obtient donc l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[X_{(1)}\alpha^{1/n}; X_{(1)}\right].$$

Pour éviter les "subtilités"de la fin, on aurait pu se douter, vu la fonction de répartition de $X_{(1)}$ que l'on allait rencontrer des problèmes et donc directement regarder des intervalles de la forme

$$IC_{1-\alpha}(\theta) = \left[X_{(1)}c_\alpha; X_{(1)}\right].$$

On résout alors

$$\begin{aligned} 1 - \alpha = \mathbb{P}\left[X_{(1)}c_\alpha \leq \theta\right] = \mathbb{P}\left[X_{(1)} \leq \frac{\theta}{c_\alpha}\right] &\Leftrightarrow \alpha = \mathbb{P}\left[X_{(1)} \geq \frac{\theta}{c_\alpha}\right] = \mathbb{P}\left[X_{(1)} - \theta \geq \frac{\theta}{c_\alpha} - \theta\right] \\ &\Leftrightarrow \alpha = \left(\frac{\theta}{\theta/c_\alpha}\right)^n = c_\alpha^n \\ &\Leftrightarrow c_\alpha = \alpha^{1/n}, \end{aligned}$$

et on retrouve ainsi l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[X_{(1)}\alpha^{1/n}; X_{(1)}\right].$$

Exemple 2 : la loi uniforme. On considère des variables aléatoires i.i.d avec $X_i \sim \mathcal{U}([0,\theta])$. On sait que l'estimateur du maximum de vraisemblance est $X_{(n)}$. Comme $X_{(n)} \leq \theta$, on va chercher un intervalle de confiance de la forme $\left[X_{(n)}, X_{(n)}c_\alpha\right]$. On a pour tout $t \in [0,\theta]$,

$$\mathbb{P}\left[X_{(n)} \leq t\right] = \mathbb{P}\left[X_1 \leq t\right]^n = \left(\frac{t}{\theta}\right)^n$$

On cherche donc c_α tel que

$$1-\alpha = \mathbb{P}\left[X_{(n)} \leq \theta \leq X_{(n)}c_\alpha\right] = \mathbb{P}\left[\theta \leq X_{(n)}c_\alpha\right] \Leftrightarrow 1-\alpha = 1-\left(\frac{1}{c_\alpha}\right)^n$$
$$\Leftrightarrow \alpha^{-1/n} = c_\alpha$$

On obtient donc l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[X_{(n)}; X_{(n)}\alpha^{-1/n}\right].$$

Remarque : On cherche souvent des intervalles de confiance tels que

$$\alpha/2 = \mathbb{P}\left[\theta \leq a\left(X_1,\ldots,X_n\right)\right] = \mathbb{P}\left[\theta \geq b\left(X_1,\ldots,X_n\right)\right].$$

Dans le cas de l'exemple 1, on a vu que pour tout $t \in (0,1)$, on a $\mathbb{P}\left[\theta \leq X_{(1)}t^{1/n}\right] = t$. Ainsi on obtient

$$\mathbb{P}\left[\theta \leq X_{(1)}\left(\frac{\alpha}{2}\right)^{1/n}\right] = \alpha/2 \qquad \text{et} \qquad \mathbb{P}\left[\theta \geq X_{(1)}\left(1-\frac{\alpha}{2}\right)^{1/n}\right] = \alpha/2.$$

On obtient donc , pour $\alpha \in (0,1)$, l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[X_{(1)}\left(\frac{\alpha}{2}\right)^{1/n}; X_{(1)}\left(1-\frac{\alpha}{2}\right)^{1/n}\right].$$

Dans le cas de l'exemple 2, on a vu que pour tout $t \in (0,1)$, on a $\mathbb{P}\left[\theta \geq X_{(n)}t^{-1/n}\right] = t$. On a donc en particulier

$$\mathbb{P}\left[\theta \geq X_{(n)}\left(\frac{\alpha}{2}\right)^{-1/n}\right] = \alpha/2 \qquad \text{et} \qquad \mathbb{P}\left[\theta \leq X_{(n)}\left(1-\frac{\alpha}{2}\right)^{-1/n}\right] = \alpha/2.$$

On obtient donc, pour $\alpha \in (0, 1/2)$, l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[X_{(n)} \left(1 - \frac{\alpha}{2}\right)^{-1/n} ; X_{(n)} \left(\frac{\alpha}{2}\right)^{-1/n} \right].$$

Il n'est malheureusement pas toujours facile de calculer la fonction de répartition de l'estimateur et ainsi construire les intervalles de confiance. On peut par contre utiliser des inégalités usuelles (Markov, Bienaymé-Tchebycheff, Chernoff...) pour obtenir des intervalles de confiances de niveau au moins $1 - \alpha$, i.e des intervalles de la forme

$$IC_{1-\alpha}(\theta) = [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]$$

avec

$$\mathbb{P}[\theta \in [a(X_1, \ldots, X_n), b(X_1, \ldots, X_n)]] \geq 1 - \alpha.$$

Exemple 1 : loi de Bernoulli. On considère des variables aléatoires i.i.d X_i suivant une loi de Bernoulli de paramètre θ. On rappelle que l'on a l'estimateur $\overline{X}_n$, et on cherche c_α tel que

$$\mathbb{P}\left[\overline{X}_n - c_\alpha \leq \theta \leq \overline{X}_n + c_\alpha\right] \geq 1 - \alpha.$$

En utilisant l'inégalité de Bienaymé-Tchebycheff, on a

$$\mathbb{P}\left[\overline{X}_n - c_\alpha \leq \theta \leq \overline{X}_n + c_\alpha\right] = \mathbb{P}\left[\left|\overline{X}_n - \theta\right| \leq c_\alpha\right] \geq 1 - \frac{\theta(1-\theta)}{nc_\alpha^2}$$

On peut alors chercher à résoudre

$$\alpha = \frac{\theta(1-\theta)}{nc_\alpha^2} \Leftrightarrow c_\alpha = \sqrt{\frac{\theta(1-\theta)}{n\alpha}}$$

et on obtiendrait l'intervalle suivant

$$IC_{1-\alpha}(\theta) = \left[\overline{X}_n \pm \sqrt{\frac{\theta(1-\theta)}{n\alpha}} \right].$$

Cependant, les bornes de l'intervalle dépendent de θ et ne sont donc pas calculables. Cependant, comme $\theta(1-\theta) \leq 1/4$, on obtient

$$\mathbb{P}\left[\left|\overline{X}_n - \theta\right| \geq c_\alpha\right] \leq \frac{1}{4nc_\alpha^2}$$

et donc l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[\overline{X}_n \pm \frac{1}{2\sqrt{n\alpha}} \right].$$

Exemple 2 : loi uniforme. On considère des variables aléatoires indépendantes et identiquement distribuées $X_1, \dots, X_n$ avec $X_1 \sim \mathcal{U}([0,\theta])$, mais on considère cette fois-ci l'estimateur obtenu via la méthode des moments, i.e $\hat{\theta}_n = 2\overline{X}_n$. On cherche c_α tel que

$$\mathbb{P}\left[\theta \in \left[\hat{\theta}_n \pm c_\alpha\right]\right]$$

ce qui revient à chercher c_α tel que

$$\mathbb{P}\left[\left|\hat{\theta}_n - \theta\right| \geq c_\alpha\right] \leq \alpha.$$

En utilisant l'inégalité de Bienaymé-Tchebycheff, on obtient

$$\mathbb{P}\left[\left|\hat{\theta}_n - \theta\right| \geq c_{alpha}\right] \leq \frac{\mathbb{V}\left[\hat{\theta}_n\right]}{c_\alpha^2}\frac{\theta^2}{3nc_\alpha^2} = \alpha \Leftrightarrow c_\alpha = \frac{\theta}{\sqrt{3n\alpha}}.$$

On obtient donc

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta}_n \pm \frac{\theta}{\sqrt{3n\alpha}}\right]$$

mais cet intervalle est inexploitable car dépendant de θ.

Remarque : On peut également chercher des intervalles de confiance de la forme

$$IC_{1-\alpha}(\theta) = \left[a\left(X1, \dots, X_n\right), +\infty\right[\qquad \text{ou} \qquad IC_{1-\alpha}(\theta) = \left]-\infty, b\left(X_1, \dots, X_n\right)\right].$$

On parle alors d'intervalles de confiance unilatères.

3.1.2 Notion de quantile

On considère une variable aléatoire X et on note F sa fonction de répartition.

Definition 3.1.1. *Pour tout $\alpha \in (0,1)$, on appelle quantile d'ordre α le réel q_α tel que*

$$q_\alpha = \inf\left\{x \in \mathbb{R}, F(x) \geq \alpha\right\}.$$

Lorsque la fonction de répartition de X est strictement croissante, alors cette dernière est bijective et le quantile q_α est alors la solution de

$$F\left(q_\alpha\right) = \alpha \Leftrightarrow q_\alpha = F^{-1}\left(\alpha\right).$$

Exemple 1 : loi uniforme. Soit $X \sim \mathcal{U}([a,b])$ et $\alpha \in (0,1)$, alors le quantile q_α d'ordre α de X est défini par

$$q_\alpha = a + (b-a)\alpha$$

En effet, on a pour tout $t \in [a,b]$, $F_X(t) = \frac{t-a}{b-a}$. On résout donc

$$\alpha = \frac{t-a}{b-a} \Leftrightarrow t = a + (b-a)\alpha.$$

Exemple 2 : loi exponentielle. Soit $X \sim \mathcal{E}(1)$ et $\alpha \in (0,1)$, alors le quantile q_α d'ordre α de X est défini par

$$q_\alpha = -\ln(1-\alpha)$$

En effet, on a pour tout $t \geq 0$, $F_X(t) = 1 - e^{-t}$ et on résout donc

$$\alpha = 1 - e^{-t} \Leftrightarrow e^{-t} = 1 - \alpha \Leftrightarrow t = -\ln(1-\alpha).$$

Exemple 3 : loi de Bernoulli. Soit $X \sim \mathcal{B}(\theta)$, alors le quantile q_α d'ordre α de X est défini par

$$q_\alpha = \begin{cases} 0 & \text{si } \alpha \in (0, 1-\theta] \\ 1 & \text{sinon} \end{cases} \tag{3.1}$$

En effet, si $\alpha \leq 1-\theta$, la plus petite valeur x telle que $F(x) \geq \alpha$ est 0 (et on a alors $F(0) = 1-\theta$). De la même façon façon, si $\alpha \in (1-\theta, 1]$, la plus petite valeur x telle que $F(x) \geq \alpha$ est 1 (et on a alors $F(1) = 1$).

Lorsque la fonction inverse est explicite (loi exponentielle, loi de Weibull, loi uniforme,...) on peut donc très facilement calculer les quantiles. Plus généralement, l'inverse de F n'est pas explicitement calculable et on utilise alors des tables statistiques pour retrouver le quantile cherché (loi normale, loi de Student, loi du Chi deux, loi de Fisher,...).

3.2 Rappels sur la loi normale

Dans ce qui suit, on considère des variables aléatoires $X_1, \ldots, X_n$ suivant des lois normales d'espérances μ_i et de variance σ_i^2. Rappelons que la fonction caractéristique d'une variable aléatoire X suivant une loi normale d'espérance μ et de variance σ^2 est définie pour tout t par

$$\Phi_X(t) = \exp\left(\mu i t - \frac{t^2\sigma^2}{2}\right).$$

Proposition 3.2.1. *Si les X_i sont indépendants, alors toute combinaison linéaire des X_i suit une loi normale. Plus précisément, soit $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$, alors*

$$\sum_{i=1}^{n} \lambda_i X_i \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

avec

$$\mu = \sum_{i=1}^{n} \lambda_i \mu_i \qquad et \qquad \sigma^2 = \sum_{i=1}^{n} \lambda_i^2 \sigma_i^2.$$

Démonstration. On note $X = \sum_{i=1}^{n} \lambda_i X_i$. Sa fonction caractéristique est définie pour tout t par (par indépendance des X_i)

$$\Phi_X(t) = \Phi_{\sum_{i=1}^{n} \lambda_i X_i}(t) = \prod_{i=1}^{n} \Phi_{\lambda_i X_i}(t) = \prod_{i=1}^{n} \Phi_{X_i}(\lambda_i t)$$

On obtient donc

$$\Phi_X(t) = \prod_{i=1}^{n} \left(\exp\left(i\mu_i\lambda_i t - \frac{\sigma_i^2 \lambda_i^2 t^2}{2} \right) \right) = \exp\left(it \sum_{i=1}^{n} \lambda_i \mu_i - \frac{t^2 \sum_{i=1}^{n} \lambda_i^2 \sigma_i^2}{2} \right),$$

qui est la fonction caractéristique d'une loi normale de paramètres $\mu = \sum_{i=1}^{n} \lambda_i \mu_i$ et $\sigma^2 = \sum_{i=1}^{n} \lambda_i^2 \sigma_i^2$. □

Definition 3.2.1 (Loi du Chi-deux)**.** *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes suivant une loi normale centrée réduite, alors la variable aléatoire*

$$Z_n = \sum_{i=1}^{n} X_i^2$$

suit une loi du Chi deux à n degrés de liberté (χ_n^2).

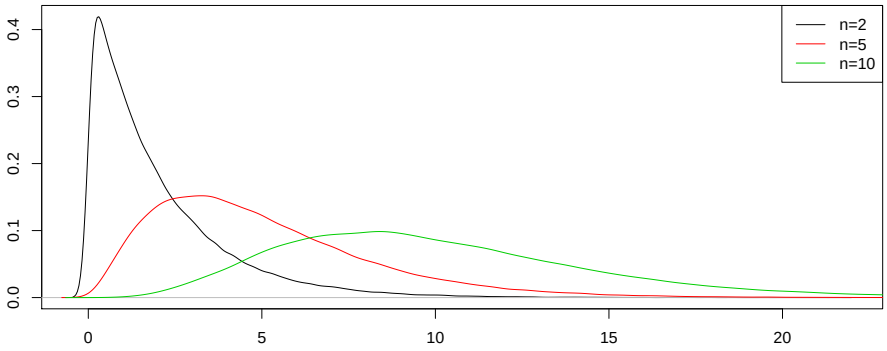

FIGURE 3.1 – Densité d'une chi deux à $n = 2, 5, 10$ degrés de liberté

Afin d'obtenir des intervalles de confiance pour l'estimateur de la moyenne, on définit maintenant la loi de Student.

Definition 3.2.2 (Loi de Student). *Soit Z, U deux variables aléatoires indépendantes telles que $Z \sim \mathcal{N}(0,1)$ et $U \sim \chi^2_n$, alors*

$$\frac{Z}{\sqrt{U/n}} \sim T_n,$$

où T_n est une loi de Student à n degrés de liberté.

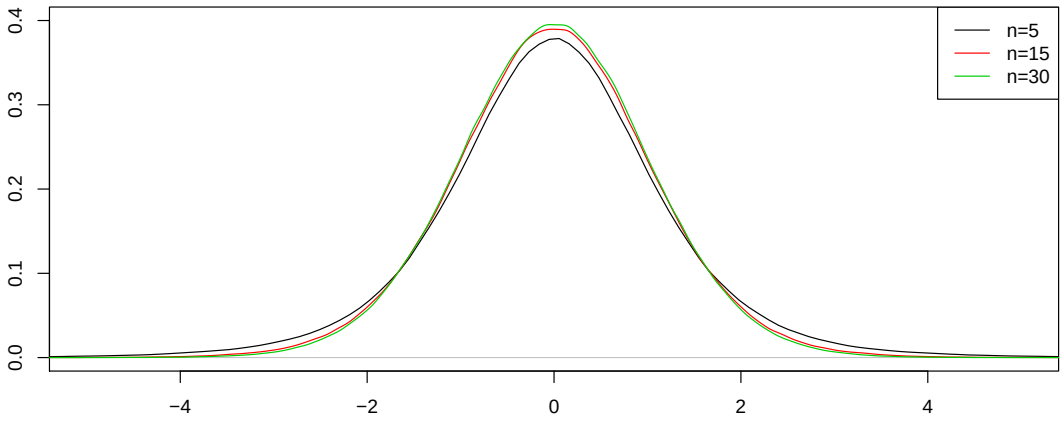

FIGURE 3.2 – Densité d'une loi de Student à $n = 5, 15, 30$ degrés de liberté.

Proposition 3.2.2. *Soit T_n une loi de Student à n degrés de liberté. Si $n \geq 2$, alors T_n admet un moment d'ordre 1 et $\mathbb{E}[T_n] = 0$. De plus, la loi de Student est symétrique en 0. Enfin,*

$$T_n \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Démonstration. A noter que si $n = 1$, on a alors, par indépendance,

$$\mathbb{E}[|T_1|] = \mathbb{E}[|Z|]\,\mathbb{E}\left[\frac{1}{|X_1|}\right] = \mathbb{E}[|Z|]\int_{\mathbb{R}} \frac{1}{|x|} e^{-\frac{x^2}{2}} dx = +\infty.$$

On admettra que pour tout $n \geq 2$, on a

$$\mathbb{E}\left[\frac{1}{\sqrt{\sum_{i=1}^n X_i^2}}\right] < +\infty.$$

Le moment d'ordre 1 de T_n existe alors (cf calculs précédents) et

$$\mathbb{E}[T_n] = \mathbb{E}[Z]\,\mathbb{E}\left[\frac{\sqrt{n}}{\sqrt{\sum_{i=1}^n X_i^2}}\right] = 0.$$

De plus, pour tout $t \in \mathbb{R}$, on a par symétrie de la loi normale

$$\mathbb{P}\left[T_n \leq t\right] = \mathbb{P}\left[\frac{Z}{\sqrt{U/n}} \leq t\right] = \mathbb{P}\left[Z \leq t\sqrt{U/n}\right] = \mathbb{P}\left[Z \geq -t\sqrt{U/n}\right] = \mathbb{P}\left[\frac{Z}{\sqrt{U/n}} \geq -t\right].$$

Enfin, par la loi des grands nombres, on a

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}\left[X_1^2\right] = 1$$

et par le théorème de continuité, on a donc

$$\frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}} \xrightarrow[n\to+\infty]{\mathbb{P}} 1.$$

Le théorème de Slutsky nous donne donc

$$\frac{Z}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

□

3.3 Intervalles de confiance dans le cas gaussien

3.3.1 Estimation de la moyenne lorsque la variance est connue

On considère dans ce qui suit des variables aléatoires $X_1, \dots X_n$ i.i.d suivant une loi normale d'espérance μ inconnue et de variance σ^2 connue.

Proposition 3.3.1. *On a*

$$\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Démonstration. Voir TD. □

Corollaire 3.3.1. *Pour tout $\alpha \in (0,1)$,*

$$\mathbb{P}\left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha,$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite, i.e si $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}\left[Z \leq q_{1-\alpha/2}\right] = 1-\alpha/2.$$

On obtient donc l'intervalle de confiance

$$IC_{1-\alpha}(\mu) = \left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Démonstration. Comme $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, en centrant et en réduisant, on obtient

$$\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \sim \mathcal{N}(0,1).$$

De plus, soit $Z \sim \mathcal{N}(0,1)$, par définition de $q_{1-\alpha/2}$, et par symétrie, on a

$$\mathbb{P}\left[Z \leq -q_{1-\alpha/2}\right] = \mathbb{P}\left[Z \geq q_{1-\alpha/2}\right] = 1 - \mathbb{P}\left[Z \leq q_{1-\alpha/2}\right] = \frac{\alpha}{2},$$

i.e $q_{\alpha/2} = -q_{1-\alpha/2}$. On a donc

$$\begin{aligned}
\mathbb{P}\left[|Z| \leq q_{1-\alpha/2}\right] &= \mathbb{P}\left[Z \leq q_{1-\alpha/2}\right] - \mathbb{P}\left[Z \leq -q_{1-\alpha/2}\right] \\
&= \mathbb{P}\left[Z \leq q_{1-\alpha/2}\right] - \mathbb{P}\left[Z \leq q_{\alpha/2}\right] \\
&= 1 - \alpha.
\end{aligned}$$

On obtient donc

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left[\frac{\sqrt{n}}{\sigma}\left|\overline{X}_n - \mu\right| \leq q_{1-\alpha/2}\right] \\
&= \mathbb{P}\left[-q_{1-\alpha/2} \leq \frac{\sqrt{n}}{\sigma}\left(\mu - \overline{X}_n\right) \leq q_{1-\alpha/2}\right] \\
&= \mathbb{P}\left[-q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu - \overline{X}_n \leq q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \\
&= \mathbb{P}\left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].
\end{aligned}$$

□

Remarque : On peut également trouver les intervalles de confiances unilatères suivant :

$$IC_{1-\alpha}(\mu) = \left[\overline{X}_n - q_{1-\alpha}\frac{\sigma}{\sqrt{n}}, +\infty\right[\qquad \text{et} \qquad IC_{1-\alpha}(\mu) = \left]-\infty, \overline{X}_n + q_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right]$$

En effet, par définition de $q_{1-\alpha}$, on a

$$1 - \alpha = \mathbb{P}\left[\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq q_{1-\alpha}\right] = \mathbb{P}\left[-\mu \leq -\overline{X}_n + \frac{\sigma}{\sqrt{n}}q_{1-\alpha}\right] = \mathbb{P}\left[\mu \geq \overline{X}_n - q_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right].$$

Pour trouver le dernier intervalle, il suffit de remarquer que $\sqrt{n}\frac{\mu - \overline{X}_n}{\sigma} \sim \mathcal{N}(0,1)$, et donc

$$1 - \alpha = \mathbb{P}\left[\sqrt{n}\frac{\mu - \overline{X}_n}{\sigma} \leq q_{1-\alpha}\right].$$

3.3.2 Estimation de la moyenne lorsque la variance est inconnue

On considère maintenant que la variance est inconnue. La proposition suivante ainsi que son corollaire sont cruciaux pour obtenir les intervalles de confiance pour la moyenne.

Proposition 3.3.2 (Admis pour le moment). *Soient* $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *et* $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2$. *Alors*

1. $\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2_{n-1}$.
2. S_n *et* $\overline{X}_n$ *sont indépendants.*

Le corollaire suivant est crucial pour obtenir des intervalles de confiance pour la moyenne dans le cas gaussien, ainsi que pour construire des tests.

Corollaire 3.3.2. *On a*

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim T_{n-1},$$

où T_{n-1} *est une loi de Student à* $n-1$ *degrés de liberté.*

Démonstration. On a la décomposition

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

et on conclut grâce à la proposition précédente. □

Le corollaire suivant donne les intervalles de confiances pour la moyenne.

Corollaire 3.3.3 (Intervalles de confiance). *Soit* $\alpha \in (0,1)$, *alors*

$$\mathbb{P}\left[\overline{X}_n - t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}\right] = 1-\alpha,$$

où $t_{n-1,1-\alpha/2}$ *est le quantile d'ordre* $1-\alpha/2$ *de la loi de Student à* $n-1$ *degrés de liberté, i.e si* $T \sim T_{n-1}$,

$$\mathbb{P}\left[T \leq t_{n-1,1-\alpha/2}\right] = 1-\alpha/2.$$

On obtient donc l'intervalle de confiance

$$IC_{1-\alpha}(\mu) = \left[\overline{X}_n - t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}; \overline{X}_n + t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}\right].$$

Démonstration. Comme la loi de Student est symétrique, on a, si $T \sim T_{n-1}$,

$$\mathbb{P}\left[|T| \leq t_{n-1,1-\alpha/2}\right] = 1-\alpha.$$

Ainsi,

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left[\left|\sqrt{n}\frac{\overline{X}_n-\mu}{S_n}\right| \leq t_{n-1,1-\alpha/2}\right] \\
&= \mathbb{P}\left[-t_{n-1,1-\alpha/2} \leq \sqrt{n}\frac{\mu-\overline{X}_n}{S_n} \leq t_{n-1,1-\alpha/2}\right] \\
&= \mathbb{P}\left[-t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu-\overline{X}_n \leq t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}\right] \\
&= \mathbb{P}\left[\overline{X}_n - t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}\right].
\end{aligned}$$

□

Remarque : Anoter que l'on peut également trouver les intervalles de confiance unilatères suivant :

$$IC_{1-\alpha} = \left[\overline{X}_n - t_{n-1,1-\alpha}\frac{X_n}{\sqrt{n}}; +\infty\right[\qquad \text{et} \qquad \left]-\infty, \overline{X}_n + t_{n-1,1-\alpha}\frac{X_n}{\sqrt{n}}\right].$$

3.3.3 Estimation de la variance

La proposition suivante nous donne les intervalles de confiance pour la variance.

Proposition 3.3.3. *Soit $\alpha \in (0,1)$. Alors*

$$\mathbb{P}\left[\frac{(n-1)S_n^2}{k_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S_n^2}{k_{\alpha/2}}\right] = 1-\alpha,$$

où $k_{\alpha/2}$ et $k_{1-\alpha/2}$ sont les quantiles d'ordres $\alpha/2$ et $1-\alpha/2$ d'une loi du chi deux à $n-1$ degrés de libertés, i.e si $Z \sim \chi^2_{n-1}$,

$$\mathbb{P}\left[Z \leq k_{\alpha/2}\right] = \alpha/2 \qquad\qquad \mathbb{P}\left[Z \leq k_{1-\alpha/2}\right] = 1-\alpha/2.$$

On obtient donc l'intervalle de confiance

$$IC_{1-\alpha}\left(\sigma^2\right) = \left[\frac{(n-1)S_n^2}{k_{1-\alpha/2}}; \frac{(n-1)S_n^2}{k_{\alpha/2}}\right].$$

Démonstration. On a vu que

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

On a donc

$$\begin{aligned}1-\alpha &= \mathbb{P}\left[k_{\alpha/2} \leq \frac{n-1}{\sigma^2} S_n^2 \leq k_{1-\alpha/2}\right] \\ &= \mathbb{P}\left[\frac{k_{\alpha/2}}{(n-1)S_n^2} \leq \frac{1}{\sigma^2} \leq \frac{k_{1-\alpha/2}}{(n-1)S_n^2}\right] \\ &= \mathbb{P}\left[\frac{(n-1)S_n^2}{k_{\alpha/2}} \geq \sigma^2 \geq \frac{(n-1)S_n^2}{k_{1-\alpha/2}}\right]\end{aligned}$$

□

3.4 Intervalles de confiance asymptotiques

Dans cette section, on s'intéresse à l'estimation du paramètre θ d'une variable aléatoire X. On considère un estimateur $\hat{\theta}_n$ asymptotiquement normal, i.e il existe $\sigma^2 > 0$ tel que

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

On supposera également que l'on a un estimateur $\hat{\sigma}_n^2$ consistant de σ^2.

Proposition 3.4.1. *Soit $\alpha \in (0,1)$,*

$$\mathbb{P}\left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + q_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha,$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

On obtient donc l'intervalle de confiance

$$IC_{1-\alpha}(\theta) = \left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}; \hat{\theta}_n + q_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right]$$

Démonstration. On a

$$Z_n := \sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} = \frac{\sigma}{\hat{\sigma}_n}\sqrt{n}\frac{\hat{\theta}_n - \theta}{\sigma},$$

et par le théorème de continuité, $\frac{\sigma}{\hat{\sigma}_n}$ converge en probabilité vers 1. On obtient, à l'aide du Théorème de Slutsky,

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

On note F_{Z_n} la fonction de répartition de Z_n et F_Z celle de Z où $Z \sim \mathcal{N}(0,1)$. Donc, pour tout x, $F_{Z_n}(x)$ converge vers $F(x)$ et en particulier

$$\mathbb{P}\left[\frac{\sqrt{n}}{\hat{\sigma}_n}\left|\hat{\theta}_n - \theta\right| \leq q_{1-\alpha/2}\right] \xrightarrow[n\to+\infty]{} \mathbb{P}\left[|Z| \leq q_{1-\alpha/2}\right] = 1-\alpha.$$

□

Remarque : Attention, ceci ne reste qu'un résultat asymptotique, et la question est de savoir à partir de quelle taille d'échantillon l'approximation par une loi normale est de bonne qualité. Si on a la chance de pouvoir avoir des intervalles non asymptotiques, il faut toujours privilégier cette option !

Exemple 1 : le lancer de pièce. On considère $X_1, \dots, X_n$ des variables aléatoires i.i.d suivant une loi de Bernoulli de paramètre $\theta \in (0,1)$. On a vu

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

On a donc

$$\mathbb{P}\left[-q_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}} \leq q_{1-\alpha/2}\right] \xrightarrow[n\to+\infty]{} 1-\alpha$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. On obtient alors

$$\mathbb{P}\left[-q_{1-\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \theta}{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}} \leq q_{1-\alpha/2}\right] \xrightarrow[n\to+\infty]{} 1-\alpha$$

$$\mathbb{P}\left[-q_{1-\alpha/2}\frac{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}}{\sqrt{n}} \leq \overline{X}_n - \theta \leq q_{1-\alpha/2}\frac{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}}{\sqrt{n}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha$$

$$\mathbb{P}\left[\overline{X}_n - q_{1-\alpha/2}\frac{\sqrt{\overline{X}_n\left(1-\overline{X}_n\right)}}{\sqrt{n}} \leq \theta \leq \overline{X}_n q_{1-\alpha/2}\frac{\sqrt{\overline{X}_n + \left(1-\overline{X}_n\right)}}{\sqrt{n}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha$$

Exemple 2 : la loi exponentielle. On considère une variable aléatoire X suivant une loi exponentielle de paramètre θ. Rappelons que l'on a alors $\mathbb{E}[X] = \theta^{-1}$ (et $\mathbb{V}[X] = \theta^{-2}$) et qu'un estimateur naturel de θ est alors $\hat{\theta}_n = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\overline{X}_n}$. On a alors, par la delta-méthode (voir TD)

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0,\theta^2\right),$$

ce que l'on peut réécrire comme

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\theta} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

A l'aide du théorème de Slutsky, on obtient

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\theta}_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

et on a donc, pour tout $\alpha \in (0,1)$

$$\mathbb{P}\left[-q_{1-\alpha/2} \leq \sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\theta}_n} \leq q_{1-\alpha/2}\right] \xrightarrow[n\to+\infty]{} 1-\alpha,$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. On obtient alors

$$\begin{aligned}
&\mathbb{P}\left[-q_{1-\alpha/2} \leq \sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\theta}_n} \leq q_{1-\alpha/2}\right] \xrightarrow[n\to+\infty]{} 1-\alpha \\
\Leftrightarrow &\mathbb{P}\left[-q_{1-\alpha/2}\frac{\hat{\theta}_n}{\sqrt{n}} \leq \hat{\theta}_n - \theta \leq q_{1-\alpha/2}\frac{\hat{\theta}_n}{\sqrt{n}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha, \\
\Leftrightarrow &\mathbb{P}\left[\hat{\theta}_n - q_{1-\alpha/2}\frac{\hat{\theta}_n}{\sqrt{n}} \leq \theta \leq \hat{\theta}_n + q_{1-\alpha/2}\frac{\hat{\theta}_n}{\sqrt{n}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha.
\end{aligned}$$

Exemple 2bis : la loi exponentielle. En réalité, on aurait pu être plus malin pour l'intervalle de confiance. En effet, le TLC peut se réécrire comme

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

En notant $q_{1-\alpha/2}$ le quantile de la loi normale centrée réduite, on obtient donc

$$\begin{aligned}
&\mathbb{P}\left[-q_{1-\alpha/2} \leq \sqrt{n}\left(\frac{\hat{\theta}_n}{\theta} - 1\right) \leq q_{1-\alpha/2}\right] \xrightarrow[n\to+\infty]{} 1-\alpha \\
&\mathbb{P}\left[1 - \frac{q_{1-\alpha/2}}{\sqrt{n}} \leq \frac{\hat{\theta}_n}{\theta} \leq 1 + \frac{q_{1-\alpha/2}}{\sqrt{n}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha \\
&\mathbb{P}\left[\frac{\hat{\theta}_n}{1 - \frac{q_{1-\alpha/2}}{\sqrt{n}}} \geq \theta \geq \frac{\hat{\theta}_n}{1 + \frac{q_{1-\alpha/2}}{\sqrt{n}}}\right] \xrightarrow[n\to+\infty]{} 1-\alpha.
\end{aligned}$$

On obtient donc l'intervalle de confiance asymptotique

$$IC_{1-\alpha}(\theta) = \left[\frac{\hat{\theta}_n}{1 + \frac{q_{1-\alpha/2}}{\sqrt{n}}}; \frac{\hat{\theta}_n}{1 - \frac{q_{1-\alpha/2}}{\sqrt{n}}}\right].$$

Chapitre 4

Vecteurs Gaussiens et théorème de Cochran

Une généralisation du Théorème Central Limite lorsque l'on doit traiter des données multivariées, i.e à valeurs dans $\mathbb{R}^d$ est la suivante :

Théorème 4.0.1 (Théorème Central Limite multivarié)**.** *Soit $X_1, \ldots, X_n$ des vecteurs aléatoires à valeurs dans $\mathbb{R}^d$, indépendants et identiquement distribués. On suppose $\mathbb{E}\left[\|X_1\|^2\right] < +\infty$ et on note $m = \mathbb{E}[X_1]$ son vecteur moyen (que l'on définira par la suite) et $\Gamma = Var[X_1]$ sa matrice de variance-covariance (que l'on définira par la suite). On note $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et on a alors la convergence en loi*

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

Ainsi, connaître et maitriser les propriétés des vecteurs gaussiens peut être crucial pour l'étude d'estimateurs lorsque le paramètre à estimer appartient à $\mathbb{R}^d$, mais pas seulement ! Sans le savoir, on a déjà vu une application du théorème de Cochran (issu de l'étude des vecteurs gaussiens que l'on verra par la suite) pour obtenir la loi de l'estimateur de la variance dans le cadre gaussien. Plus précisément, dans ce cours (et en TD), on utilisera les vecteurs gaussiens pour les problèmes suivants :

- Obtention d'intervalles de confiance non asymptotique de la moyenne et de la variance d'une variable aléatoire suivant une loi normale (cf chapitre précédent).
- Estimation des paramètres dans le cadre du modèle linéaire gaussien, i.e on considère :
  - des variables aléatoires $Y_1, \ldots, Y_n$ que lon cherche à expliquer
  - pour tout i, des variables explicatives $x_{i,1}, \ldots, x_{i,p}$
  - des variables aléatoires indépendantes $\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$
  - le modèle de régression
  $$Y_i = a_1 x_{i,1} + \ldots + a_p x_{i,p} + \epsilon_i$$
  et l'objectif est donc d'estimer $a_1, \ldots, a_p$, construire des intervalles de confiance... (voir TD).
- Construction de tests dans le cadre gaussien. On verra par la suite comment tester la va-

leur d'une moyenne, d'une variance, comparer les moyennes de différentes variables aléatoires....

4.1 Vecteurs aléatoires

Soit $X = \begin{pmatrix} X_1 \\ \dots \\ X_d \end{pmatrix} \in \mathbb{R}^d$ un vecteur aléatoire tel que $\mathbb{E}\left[X_j^2\right] < +\infty$. Pour tout i, j, la covariance entre X_i et X_j est définie par

$$\mathrm{Cov}\left(X_i, X_j\right) = \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(X_j - \mathbb{E}\left[X_j\right]\right)\right] = \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right].$$

Remarque : Si X_i et X_j sont indépendantes, $\mathrm{Cov}\left(X_i, X_j\right) = 0$ mais la réciproque est généralement fausse (sauf pour les vecteurs gaussiens, ce que l'on verra par la suite).

Definition 4.1.1. *Soit X un vecteur aléatoire de $\mathbb{R}^d$ tel que pour tout $i = 1, ..., d$, $\mathbb{E}\left[X_i^2\right] < +\infty$. Le vecteur moyen de X est défini par*

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

et sa matrice de covariance est définie par

$$\begin{aligned} Var\left[X\right] &= \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)\left(X - \mathbb{E}\left[X\right]\right)^T\right] \\ &= \begin{pmatrix} Var\left[X_1\right] & Cov\left(X_1, X_2\right) & \dots & Cov\left(X_1, X_d\right) \\ Cov\left(X_2, X_1\right) & Var\left[X_2\right] & \dots & Cov\left(X_2, X_d\right) \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(X_d, X_1\right) & Cov\left(X_d, X_2\right) & \dots & Var\left[X_d\right] \end{pmatrix} \end{aligned}$$

A noter que pour que le vecteur moyen soit défini, il suffit de supposer que pour tout i, la variable aléatoire X_i est intégrable.

Théorème 4.1.1. *La matrice de covariance est une matrice symétrique semi-définie positive.*

La symétrie vient de la symétrie de la covariance, i.e pour tout $i, ; j$,

$$\mathrm{Cov}\left(X_i, X_j\right) = \mathrm{Cov}\left(X_j, X_i\right).$$

De plus, on a pour tout $v = \left(v_1, \dots, v_d\right)^T$

$$v^T \mathrm{Var}(X) = \left(\sum_{i=1}^{d} v_i \mathrm{Cov}\left(X_i, X_1\right), \sum_{i=1}^{d} v_i \mathrm{Cov}\left(X_i, X_2\right), \dots, \sum_{i=1}^{d} v_i \mathrm{Cov}\left(X_i, X_d\right)\right)$$

et on obtient donc

$$\begin{aligned} v^T \mathrm{Var}(X) v &= \sum_{j=1}^{d} \sum_{i=1}^{d} v_i \mathrm{Cov}\left(X_i, X_j\right) v_j \\ &= \sum_{j=1}^{d} \sum_{i=1}^{d} \mathrm{Cov}\left(v_i X_i, v_j X_j\right) \\ &= \sum_{i=1}^{d} \mathbb{V}\left[v_i X_i\right] + \sum_{i \neq j} \mathrm{Cov}\left(v_i X_i, v_j X_j\right) \\ &= \mathrm{Var}\left(v_1 X_1 + \ldots + v_d X_d\right) \geq 0. \end{aligned}$$

Théorème 4.1.2. *Soit X un vecteur aléatoire de $\mathbb{R}^d$ de moyenne m et de matrice de covariance C. Alors, si A est une matrice réelle $p \times d$, le vecteur aléatoire $AX \in \mathbb{R}^p$ a pour vecteur moyen Am et pour matrice de covariance ACA^T.*

Démonstration. Par linéarité de l'espérance, on a

$$\mathbb{E}[AX] = A\mathbb{E}[X] = Am$$

Si vraiment on veut s'en convaincre, on a

$$AX = \left(\sum_{i=1}^{d} a_{1,j} X_j, \ldots, \sum_{i=1}^{d} a_{p,j} X_j\right) \quad \text{et} \quad Am = \left(\sum_{i=1}^{d} a_{1,j} \mathbb{E}\left[X_j\right], \ldots, \sum_{i=1}^{d} a_{p,j} \mathbb{E}\left[X_j\right]\right)$$

et par définition de l'espérance multidimensionnelle et par linéarité, on obtient le résultat. □

Definition 4.1.2. *Soit X un vecteur aléatoire de $\mathbb{R}^d$. Sa fonction caractéristique Φ_X est définie pour tout $u \in \mathbb{R}^d$ par*

$$\Phi_X(u) = \mathbb{E}\left[e^{iu^T X}\right] = \mathbb{E}\left[e^{i(u_1 X_1 + \ldots + u_d X_d)}\right].$$

4.2 Vecteurs aléatoires gaussiens

Definition 4.2.1. *Soit X un vecteur aléatoire de $\mathbb{R}^d$. X est un vecteur gaussien de $\mathbb{R}^d$ si et seulement si toute combinaison linéaire de ses composantes est une variable aléatoire réeelle gaussienne, i.e*

$$\forall u \in \mathbb{R}^d, \quad u^T X \sim \mathcal{N}\left(m_u, \sigma_u^2\right).$$

Exemple : Un vecteur X dont les composantes X_i sont indépendantes et suivent une loi normale est un vecteur gaussien.

Théorème 4.2.1. *Soit X un vecteur aléatoire de $\mathbb{R}^d$. Les assertions suivantes sont équivalentes :*

1. *Le vecteur X est un vecteur gaussien de moyenne m et de matrice de covariance Γ.*
2. *La fonction caractéristique de X est donnée pour tout $u \in \mathbb{R}^d$ par*

$$\Phi_X(u) = \mathbb{E}\left[e^{iu^T X}\right] = \exp\left(iu^T m - \frac{1}{2}u^t \Gamma u\right).$$

Démonstration. Soit $X \sim \mathcal{N}(m, \Gamma)$. A noter que pour tout $u \in \mathbb{R}^d$, on a par définition $u^T X$ qui suit une loi normale et plus précisément

$$u^T X \sim \mathcal{N}\left(u^T m, u^T \Gamma u\right).$$

On a donc

$$\Phi_{u^T X}(t) = \mathbb{E}\left[e^{itu^T X}\right] = \exp\left(itu^T m - \frac{1}{2}u^T \Gamma u t^2\right).$$

On obtient donc pour tout u,

$$\Phi_X(u) = \mathbb{E}\left[e^{iu^T X}\right] = \Phi_{u^T X}(1) = \exp\left(iu^T m - \frac{1}{2}u^T \Gamma u\right).$$

□

Corollaire 4.2.1. *Soit $X \sim \mathcal{N}(m, \Gamma)$ à valeurs dans $\mathbb{R}^d$. Soit A une matrice $p \times d$ et $b \in \mathbb{R}^p$, alors*

$$AX + b \sim \mathcal{N}\left(Am + b, A\Gamma A^T\right).$$

Démonstration. On a pour tout $u \in \mathbb{R}^p$,

$$\begin{aligned}
\Phi_{AX+b}(u) &= \mathbb{E}\left[\exp\left(iu^T AX + u^T b\right)\right] \\
&= \exp\left(iu^T b\right) \mathbb{E}\left[\exp\left(iu^T AX\right)\right] \\
&= \exp\left(iu^T b\right) \Phi_X\left(A^T u\right) \\
&= \exp\left(iu^T b\right) \exp\left(iu^T Am - \frac{1}{2}u^T A\Gamma A^T u\right) \\
&= \exp\left(iu^T (Am + b) - \frac{1}{2}u^T \left(A\Gamma A^T\right) u\right).
\end{aligned}$$

□

Corollaire 4.2.2. *Soit $X \sim \mathcal{N}(m, \Gamma)$, si Γ est inversible, alors*

$$\Gamma^{-1/2}(X - m) \sim \mathcal{N}(0, I_d).$$

Démonstration. Soit $Y = \Gamma^{-1/2}(X - m)$, d'après le corollaire précédent, Y est un vecteur gaussien

et par linéarité

$$\mathbb{E}[Y] = \Gamma^{-1/2}\mathbb{E}\left[X - m\right] = 0.$$

De plus, comme $Y = \Gamma^{-1/2}X - \Gamma^{-1/2}m$,

$$\mathrm{Var}\left[Y\right] = \Gamma^{-1/2}\Gamma\left(\Gamma^{-1/2}\right)^T = I_d.$$

□

Théorème 4.2.2. *Soit $X \in \mathbb{R}^d$ avec $X \sim \mathcal{N}(m, \Gamma)$. X admet une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$ si et seulement si Γ est inversible. Dans ce cas, la densité de f est donnée pour tout $x \in \mathbb{R}^d$ par*

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \left|det(\Gamma)\right|}} \exp\left(-\frac{1}{2}(x-m)^T\Gamma^{-1}(x-m)\right)$$

Théorème 4.2.3. *Pour tout vecteur gaussien X de $\mathbb{R}^d$, les propriétés suivantes sont équivalentes :*

- *Les composantes $X_1, \ldots, X_d$ de X sont indépendantes.*
- *La matrice de covariance Γ de X est diagonale.*

Démonstration. l'implication $(1) \Rightarrow (2)$ est évidente. Dans l'autre sens, on rappelle que des variables aléatoires $Y_1, ..., Y_d$ sont indépendantes si et seulement si pour tout $u \in \mathbb{R}^d$,

$$\Phi_{(Y_1,\ldots,Y_d)}(u) = \prod_{i=1}^{d} \Phi_{Y_i}\left(u_i\right)$$

On note

$$\Gamma = \mathrm{Diag}\left(\sigma_1^2, \ldots, \sigma_d^2\right) = \mathrm{Diag}\left(\mathrm{Var}\left(X_1\right), \ldots, \mathrm{Var}\left(X_d\right)\right)$$

la matrice de covariance de X. On a alors pour tout $u \in \mathbb{R}^d$,

$$\begin{aligned}
\Phi_X(u) &= \exp\left(iu^Tm - \frac{1}{2}u^T\Gamma u\right) \\
&= \exp\left(i\sum_{j=1}^{d} u_j m_j - \frac{1}{2}\sum_{j=1}^{d}\sigma_j^2 u_j^2\right) \\
&= \prod_{j=1}^{d} \exp\left(iu_j m_j - \frac{1}{2}\sigma_j^2 u_j^2\right) \\
&= \prod_{j=1}^{d} \Phi_{X_j}\left(u_j\right).
\end{aligned}$$

□

Exemple : Soit $X = \left(X_1, X_2\right)^T \sim \mathcal{N}\left(0, V\right)$. Alors la variable aléatoire $Z = X_1 + X_2 \sim \mathcal{N}\left(0, u^TVu\right)$. En effet il suffit juste de remarquer que $Z = u^TX$.

Corollaire 4.2.3. *Soit X un vecteur gaussien, alors pour tout $i \neq j$, X_i et X_j sont indépendantes si et seulement si*

$$Cov(X_i, X_j) = 0.$$

Démonstration. La première implication est claire. Pour tout i, j tel que $i \neq j$, on note $\tilde{X} = \left(X_i, X_j\right)^T$ qui est un vecteur gaussien de matrice de covariance

$$\mathrm{Var}\left(\tilde{X}\right) = \begin{pmatrix} \mathrm{Var}\left(X_i\right) & \mathrm{Cov}\left(X_i, X_j\right) \\ \mathrm{Cov}\left(X_i, X_j\right) & \mathrm{Var}\left(X_j\right) \end{pmatrix}$$

Si $\mathrm{Cov}\left(X_i, X_j\right) = 0$ la matrice est alors diagonale et X_i, X_j sont donc indépendants. □

Attention ! Encore une fois, ceci est vrai car X est un vecteur gaussien ! Il n'y a pas d'implication générale !

Contre-exemple : Soit $X \sim \mathcal{N}(0,1)$ et Y une variable aléatoire suivant une loi de Rademacher de paramètre $p \in (0,1)$, i.e

$$\mathbb{P}\left[Y = 1\right] = p \qquad \text{et} \qquad \mathbb{P}\left[Y = -1\right] = 1 - p.$$

Dans ce qui suit, on prend $p = 1/2$, on considère que X et Y sont indépendants, et on considère les variables aléatoires X et $Z = XY$. A noter que $\mathbb{E}[Y] = 0$, et donc $\mathbb{E}\left[Z\right] = 0$. De plus

$$\mathrm{Cov}\left(X, Z\right) = \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)\left(Z - \mathbb{E}[Z]\right)\right] = \mathbb{E}\left[XZ\right] = \mathbb{E}\left[X^2 Y\right] = \mathbb{E}\left[X^2\right]\mathbb{E}[Y] = 0.$$

Cependant, on a

$$\begin{aligned} \mathbb{P}\left[X \leq 1, Z \leq -2\right] = \mathbb{P}\left[X \leq 1, XY \leq -2\right] \quad &= \mathbb{P}\left[X \leq 1, X \leq -2, Y = 1\right] + \underbrace{\mathbb{P}\left[X \leq 1, -X \leq -2, Y = -1\right]}_{=0} \\ &= \frac{1}{2}\mathbb{P}\left[X \leq -2\right] \end{aligned}$$

et par symétrie de la loi normale

$$\begin{aligned} \mathbb{P}\left[X \leq 1\right]\mathbb{P}\left[Z \leq -2\right] &= \mathbb{P}\left[X \leq 1\right]\left(\mathbb{P}\left[X \leq -2, Y = 1\right] + \mathbb{P}\left[-X \leq -2, Y = -1\right]\right) \\ &= \mathbb{P}\left[X \leq 1\right]\left(\frac{1}{2}\mathbb{P}\left[X \leq -2\right] + \frac{1}{2}\mathbb{P}\left[-X \leq -2\right]\right) \\ &= \underbrace{\mathbb{P}\left[X \leq 1\right]}_{\neq 1/2}\mathbb{P}\left[X \leq -2\right] \\ &\neq \mathbb{P}\left[X \leq 1, Z \leq -2\right]. \end{aligned}$$

Les variables X et Z ne sont donc pas indépendantes. On obtient donc un exemple de variables aléatoires avec une covariance nulle mais pas indépendante. On peut aussi en conclure que le

vecteur $(X, Z)^T$ n'est pas gaussien. Or pour tout $x \in \mathbb{R}$, par symétrie de la loi normale,

$$\mathbb{P}[Z \leq x] = \mathbb{P}[X \leq x, Y = 1] + \mathbb{P}[-X \leq x, Y = -1] = \frac{1}{2}(\mathbb{P}[X \leq x] + \mathbb{P}[X \geq -x]) = \mathbb{P}[X \leq x].$$

Z suit donc la même loi que X, i.e Z suit une loi normale. On a donc un exemple de vecteur $((X, Z)^T)$ dont les coordonnées suivent des loi normales mais qui n'est pas gaussien.

4.3 Théorème de Cochran et applications

Théorème 4.3.1 (Théorème de Cochran)**.** *Soit X un vecteur aléatoire de $\mathbb{R}^n$ suivant une loi $\mathcal{N}(m, \sigma^2 I_n)$ avec $\sigma^2 > 0$. Soit $\mathbb{R}^n = E_1 \oplus E_2 \oplus \ldots \oplus E_p$ une décomposition de $\mathbb{R}^n$ en somme directe de p sous-espaces vectoriels orthogonaux de dimensions respectives $d_1, \ldots, d_p$. Soit P_k le projecteur orthogonal sur E_k et $Y_k = P_k X$ la projection orthogonale de X sur E_k.*

1. *Les projections $Y_1, \ldots, Y_p$ sont des vecteurs gaussiens indépendants et $Y_k \sim \mathcal{N}(P_k m, \sigma^2 P_k)$.*
2. *Les variables aléatoires $\|Y_1 - P_1 m\|^2, \ldots, \|Y_p - P_p m\|^2$ sont indépendantes et*

$$\frac{\|Y_k - P_k m\|^2}{\sigma^2} \sim \chi^2_{d_k}$$

Remarque : A noter que pour tout $k \neq k'$, on a également indépendance entre Y_k et $\|Y_{k'} - P_{k'} m\|^2$.

Démonstration. On note

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} P_1 \\ \vdots \\ P_p \end{pmatrix} X = AX$$

On a donc $Y \sim \mathcal{N}(Am, \sigma^2 AA^T)$. De plus comme les P_k sont des projecteurs orthogonaux, on a $P_k^2 = P_k$ et $P_k P_{k'} = 0$ si $k \neq k'$. On a donc $AA^T = \text{Diag}(P_1, \ldots, P_k)$ qui est diagonale par block et on a donc que les Y_k sont indépendants, et en particulier les variables aléatoires $\|Y_k - P_k m\|^2$ sont indépendantes. De plus

$$Y_k \sim \mathcal{N}(P_k m, \sigma^2 P_k).$$

Soit $B_k = \{e_{k,1}, \ldots, e_{k,d_k}\}$ une base orthonormée de E_k, on a alors

$$Y_k - P_k m = \sum_{j=1}^{d_k} \langle X - m, e_{k,j} \rangle e_{k,j},$$

et en complétant la base B_k en une base orthonormée de $\mathbb{R}^n$, on a $Y_k - P_k m$ et sa matrice de cova-

riance qui s'écrivent

$$(Y_k - P_k m)_{B_k} = \begin{pmatrix} \langle X - m, e_{k,1} \rangle \\ \vdots \\ \langle X - m, e_{k,d_k} \rangle \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \text{et} \qquad \operatorname{Var}(Y_k - P_k m)_{B_k} = \begin{pmatrix} \sigma^2 I_{d_k} & 0 \\ 0 & 0 \end{pmatrix}$$

et on on a donc que les $\langle X - m, e_{k,j} \rangle$ sont indépendants et pour $j = 1, \ldots, d_k$,

$$\langle X - m, e_{k,j} \rangle \sim \mathcal{N}(0, \sigma^2)$$

et donc

$$\frac{1}{\sigma^2} \|Y_k - P_k m\|^2 = \sum_{j=1}^{d_k} \frac{\langle X - m, e_{k,j} \rangle^2}{\sigma^2} \sim \chi^2_{d_k}.$$

□

Corollaire 4.3.1. *Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées avec $X_1 \sim \mathcal{N}(m, \sigma^2)$. On a alors*

1. $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2_{n-1}$.
2. $\overline{X}_n$ *et* S_n^2 *sont indépendantes.*

Démonstration. On note $X = (X_1, \ldots, X_n)^T \sim \mathcal{N}(me, \sigma^2 I_n)$ avec $e = (1, \ldots, 1)^T \in \mathbb{R}^n$ et $u = \frac{e}{\|e\|} = n^{-1/2}(1, \ldots, 1)$. On note $E = \operatorname{Vect}\{u\}$ et

$$\mathbb{R}^n = E \oplus E^\perp.$$

On note P_E la projection orthogonale sur E, et on a

$$P_E X = \langle X, u \rangle u = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i u = \overline{X}_n e.$$

De plus, $P_E(me) = me$ et $P_{E^\perp}(me) = 0$. De plus,

$$P_{E^\perp}(X) = (I_n - P_E)(X) = X - \overline{X}_n e = \begin{pmatrix} X_1 - \overline{X}_n \\ \vdots \\ X_n - \overline{X}_n \end{pmatrix}.$$

Enfin,

$$\|P_E(X) - P_E(me)\|^2 = n(\overline{X}_n - m)^2$$

et

$$\|P_{E^\perp}(X) - P_{E^\perp}(me)\|^2 = \sum_{i=1}^{n} (X_i - m)^2 = (n-1)S_n^2.$$

On a donc, comme $\dim(E) = 1$, en appliquant le théorème de Cochran,

1. $n\left(\overline{X}_n - m\right)^2$ et $(n-1)S_n^2$ sont indépendantes, i.e $\overline{X}_n$ et S_n^2 sont indépendantes.
2. On a

$$\frac{n\left(\overline{X}_n - m\right)^2}{\sigma^2} \sim \chi_1^2 \qquad \text{et} \qquad \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2.$$

□

Proposition 4.3.1. *Soient Z, Z_p, Z_q trois variables aléatoires telles que*

1. $Z = Z_p + Z_q$
2. $Z_p \sim \chi_p^2$ *et* $Z_q \sim \chi_q^2$
3. Z_p *et* Z_q *sont indépendantes*

alors $Z \sim \chi_{p+q}^2$.

Démonstration. Soit $Z'_p = \sum_{i=1}^{p} X_i^2$ et $Z'_q = \sum_{i=p+1}^{p+q} X_i^2$ où les X_i sont indépendants et $X_i \sim \mathcal{N}(0,1)$. On a Z'_p et Z'_q qui sont indépendants et pour tout t,

$$\Phi_{Z'_p+Z'_q}(t) = \Phi_{Z'_p}(t)\Phi_{Z'_q}(t) = \Phi_{Z_p}(t)\Phi_{Z_q}(t) = \Phi_Z(t).$$

et donc Z et $Z'_p + Z'_q$ ont la même loi. Or

$$Z'_p + Z'_q = \sum_{i=1}^{p+q} X_i^2 \sim \chi_{p+q}^2.$$

□

Chapitre 5

Tests

Un test statistique est une procédure mathématique qui permet, à l'aide d'un modèle probabiliste, de rejeter ou non une hypothèse avec un certain risque connu. En pratique, on cherche à expliquer les différences observées entre une hypothèse posée a priori et le résultat obtenu à partir des données. Le test permet alors de décider si ces différences sont seulement dues à la fluctuation d'échantillonnage ou bien si elles sont significatives. Cette dernière possibilité impliquerait alors que la ou les hypothèses posées a priori ne sont pas fondées.

5.1 Généralités

5.1.1 Généralités

Le principe d'un test consiste à

1. Poser une hypothèse $H0$ et une hypothèse alternative $H1$
2. A l'aide du modèle probabiliste posé, définir une zone de rejet de l'hypothèse nulle $H0$, i.e une zone où les résultats expérimentaux seraient trop éloignés du résultat théorique pour être crédibles.
3. Regarder si le résultat expérimental est dans cette zone de rejet.
4. Si le résultat est dans la zone de rejet, on rejette l'hypothèse nulle $H0$.
5. Si le résultat n'est pas dans la zone de rejet, on ne rejette pas $H0$.

Attention ! La sémantique est importante, on dit qu'on ne rejette pas $H0$ et non pas que l'on accepte $H0$. Même si la différence semble faible entre ces deux expressions, on verra par la suite pourquoi c'est important.

Exemple : On considère l'exemple du lancer de pièce. On jette 10 fois une pièce et on obtient 9 "Face". Un test statistique nous permettra de dire avec un certain risque si cette proportion de "Face" est seulement due à la fluctuation d'échantillonnage ou bien si la pièce est truquée.

5.1.2 Etapes d'un test statistique

On dispose de n données $x_1, \ldots, x_n$ qui sont des mesures de variables qualitatives ou quantitatives.
Introduire un modèle probabiliste : On peut par exemple faire l'hypothèse que les x_i sont des réalisations de n variables aléatoires indépendantes et identiquement distribuées $X_1, \ldots, X_n$.

Dans le cas du lancer de pièce, on notera $x_i = 1$ si le i-ème lancer donne "Pile" et 0 sinon. On considère alors les x_i comme étant les réalisations de v.a i.i.d X_i suivant une loi de Bernoulli de paramètre θ.

Les hypothèses nulle et alternative : Disposant d'un modèle probabiliste, on souhaite vérifier une hypothèse sur ce modèle. On appelle cette hypothèse l'hypothèse nulle $H0$, et on introduit sa négation (hypothèse alternative) $H1$.

Dans le cas du lancer de pièce on veut savoir si la pièce est truquée ou non, i.e si $\theta = 0.5$. On pose alors les hypothèses :

$$H0 : "\theta = 0.5" \qquad \text{contre} \qquad H1 : "\theta \neq 0.5"$$

On peut également se demander si la probabilité d'obtenir "Pile" n'est pas en dessous de 0.5 et on pose alors

$$H0 : "\theta < 0.5" \qquad \text{contre} \qquad H1 : "\theta \geq 0.5"$$

La statistique de test et sa loi sous $H0$: Une fois que l'on a introduit le modèle probabiliste et les hypothèses, il faut pouvoir disposer d'une variable aléatoire appelée statistique de test, que l'on notera Z et :

— On connaît sa loi sous $H0$.
— Sa loi n'est pas la même sous $H1$.

Bien évidemment, Z est définie en fonction des variables aléatoires $X_1, ..., X_n$, i.e $Z = T(X_1, \ldots X_n)$ et si on note $z_{obs} = T(x_1, \ldots, x_n)$, z_{obs} doit être calculable.

Zone de rejet : Connaissant la loi de Z sous $H0$, on peut déterminer ses valeurs les plus extrêmes, i.e déterminer une zone ZR où $\mathbb{P}[Z \in ZR] \leq \alpha$ si $H0$ est vraie. Le seuil α est choisi librement. Dans le cas où on teste :

1. $H0 : "\theta = \theta_0"$, les valeurs "extrêmes" sont les valeurs les plus éloignées de la valeur attendue.
2. $H0 : "\theta < \theta_0"$, les valeurs "extrêmes" sont les grandes valeurs.
3. $H0 : "\theta > \theta_0"$, les valeurs "extrêmes" sont les petites valeurs.

Conclusion : On conclut le test à partir de la valeur observée z_{obs} :

— Si z_{obs} appartient à la zone de rejet, on rejette $H0$ au risque α.
— Si z_{obs} n'appartient pas à la zone de rejet, on ne rejette pas $H0$ au risque α.

Si on reprend l'exemple du lancer de pièce, on souhaite tester si la pièce est truquée ou non au risque de 5%, i.e on teste

$$H0 : "\theta = 0.5" \qquad \text{contre} \qquad H1 : "\theta \neq 0.5"$$

Sous $H0$, X_i suit une loi de Bernoulli de paramètre 0.5 et $Z = \sum_{i=1}^{10} X_i$ suit une loi binomiale de paramètre $n = 10$, $\theta = 0.5$ sous H0, ce qui est faux sous $H1$. On a donc notre statistique de test. Comme les quantiles d'ordres $\alpha/2$ et $1 - \alpha/2$ de la loi binomiale de paramètres $10, 0.5$ sont 2 et 8, on a la zone de rejet

$$ZR_{H0} = \{0, 1\} \cup \{9, 10\}$$

Nous avons observé 9 "Face", donc $z_{obs} \in ZR_{H0}$ et on rejette donc $H0$ au risque de 5%, i.e au risque de 5%, on peut dire que la pièce est truquée.

p-value : Plus le seuil α est petit, plus le test est fiable. Comme il est inconcevable de faire les tests pour toutes les valeurs α possibles, on va s'intéresser à la p-value qui est définie comme

$$p - \text{value} = \inf \{\alpha \in (0, 1),\ "\text{On rejette H0}"\}.$$

On verra par la suite comment calculer la p-value pour les différents tests. La p-value peut donc être vue comme étant, sous $H0$, la probabilité d'avoir eu un "aussi mauvais résultat" que z_{obs}.

— Si $p - \text{value} \geq \alpha$, alors on ne rejette pas $H0$

— Si $p - \text{value} < \alpha$, alors on rejette $H0$

Le risque α peut donc être vu comme étant un seuil subjectif où on estime qu'un évènement ayant une probabilité inférieure à ce seuil de se réaliser ne devrait pas se réaliser.

5.2 Tests sur la moyenne et la variance

On considère maintenant que les données réelles $x_1, \ldots, x_n$ sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ avec $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ et μ, σ^2 sont inconnus.

5.2.1 Test de conformité d'une moyenne

On veut savoir avec un risque α si la moyenne théorique μ est différente ou non d'une valeur μ_0 donnée.

Construction de la statistique de test : On dispose de l'estimateur $\overline{X}_n$ de la moyenne μ qui vérifie

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Comme la variance σ^2 est inconnue, on utilise alors

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim T_{n-1},$$

où T_{n-1} suit une loi de Student à $n-1$ degrés de liberté. Enfin, sous H0, comme $\mu = \mu_0$, on a alors

$$\sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \sim T_{n-1}.$$

ce qui est bien évidemment faux sous H1. En effet, sous $H1$,

$$\sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} = \sqrt{n}\frac{\overline{X}_n - \mu}{S_n} + \sqrt{n}\frac{\mu - \mu_0}{S_n}$$

et ne suit donc pas une loi de Student à $n-1$ degrés de liberté.

Le test : On teste au risque α l'hypothèse nulle $H0$: "$\mu = \mu_0$" contre l'hypothèse alternative $H1$: "$\mu \neq \mu_0$", on utilise la statistique de test

$$Z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \sim T_{n-1}, \qquad \text{sous } H0.$$

On définit la zone de rejet par

$$ZR = \{|Z| > t_{n-1,1-\alpha/2}\}$$

où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté. On calcule alors $z_{obs} = \sqrt{n}\frac{\overline{x}_n - \mu_0}{s_n}$.
Si $|z_{obs}| > t_{n-1,1-\alpha/2}$, on rejette $H0$, sinon, on ne rejette pas $H0$.

Remarque 1 : Notons que faire ce test revient donc à vérifier que μ_0 appartient à la réalisation de l'intervalle de confiance de niveau $1-\alpha$ obtenu à partir des données observées. Si μ_0 appartient à cet intervalle, alors on ne rejette pas $H0$ et inversement. En effet, on a

$$\begin{aligned}
\text{"On ne rejette pas H0"} &\Leftrightarrow |z_{obs}| \leq t_{n-1,1-\alpha/2} \\
&\Leftrightarrow -t_{n-1,1-\alpha/2} \leq z_{obs} \leq t_{n-1,1-\alpha/2} \\
&\Leftrightarrow t_{n-1,1-\alpha/2} \leq \sqrt{n}\frac{\overline{x}_n - \mu_0}{s_n} \leq t_{n-1,1-\alpha/2} \\
&\Leftrightarrow -t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}} \leq \overline{x}_n - \mu_0 \leq t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}} \\
&\Leftrightarrow -\overline{x}_n - t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}} \leq -\mu_0 \leq -\overline{x}_n + t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}} \\
&\Leftrightarrow \overline{x}_n - t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}} \leq \mu_0 \leq \overline{x}_n + t_{n-1,1-\alpha/2}\frac{s_n}{\sqrt{n}} \\
&\Leftrightarrow \mu_0 \in IC_{1-\alpha}(\mu).
\end{aligned}$$

Remarque 2 : En réalité, pour trouver la zone de rejet, comme $\overline{X}_n$ doit être proche de μ_0, on cherche c_α tel que sous H0

$$\mathbb{P}_{\mu=\mu_0}\left["\text{On rejette H0"}\right] = \mathbb{P}_{\mu=\mu_0}\left[\left|\overline{X}_n - \mu_0\right| \geq c_\alpha\right] = \alpha.$$

On obtient alors

$$\begin{aligned}\alpha &= \mathbb{P}_{\mu=\mu_0}\left[\left|\overline{X}_n - \mu_0\right| \geq c_\alpha\right] \\ &= \mathbb{P}_{\mu=\mu_0}\left[\sqrt{n}\frac{\left|\overline{X}_n - \mu_0\right|}{S_n} \geq \sqrt{n}\frac{c_\alpha}{S_n}\right] \\ &= \mathbb{P}_{\mu=\mu_0}\left[\sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \leq -\sqrt{n}\frac{c_\alpha}{S_n}\right] + \mathbb{P}_{\mu=\mu_0}\left[\sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \geq \sqrt{n}\frac{c_\alpha}{S_n}\right].\end{aligned}$$

Or si on note F_T la fonction de répartition de la loi de Student à $n-1$ degrés de liberté, on obtient par symétrie de la loi de Student,

$$\begin{aligned}\alpha &= F_T\left(-\sqrt{n}\frac{c_\alpha}{S_n}\right) + 1 - F_T\left(\sqrt{n}\frac{c_\alpha}{S_n}\right) \\ &= 2 - 2F_T\left(\sqrt{n}\frac{c_\alpha}{S_n}\right)\end{aligned}$$

et donc

$$F_T\left(\sqrt{n}\frac{c_\alpha}{S_n}\right) = 1 - \alpha/2.$$

Ainsi, en appliquant F_T^{-1},

$$t_{n-1,1-\alpha/2} = \sqrt{n}\frac{c_\alpha}{S_n},$$

et donc

$$c_\alpha = t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}.$$

On retrouve alors

$$ZR = \left\{\left|\overline{X}_n - \mu_0\right| \geq t_{n-1,1-\alpha/2}\frac{S_n}{\sqrt{n}}\right\} = \left\{\left|\sqrt{n}\frac{X_n - \mu_0}{S_n}\right| \geq t_{n-1,1-\alpha/2}\right\}.$$

Remarque 3 : Soit $T \sim T_{n-1}$, alors on peut calculer la p-value comme

$$p - \text{value} = \mathbb{P}\left[|T| > |z_{obs}|\right].$$

En effet, on a

$$\begin{aligned}p - \text{value} &= \inf\{\alpha \in (0,1), \text{"On rejette H0"}\} \\ &= \inf\{\alpha \in (0,1), |z_{obs}| \geq t_{n-1,1-\alpha/2}\} \\ &= \inf\{\alpha \in (0,1), F_T(|z_{obs}|) \geq 1 - \alpha/2\} \\ &= \inf\{\alpha \in (0,1), \alpha \geq 2 - 2F_T(|z_{obs}|)\} \\ &= 2 - 2F_T(|z_{obs}|) \\ &= 2\mathbb{P}\left[T \geq |z_{obs}|\right]\end{aligned}$$

où $T \sim T_{n-1}$. Par symétrie de la loi de Student,

$$p - \text{value} = 2\mathbb{P}\left[T \geq |z_{obs}|\right] = \mathbb{P}\left[T \geq |z_{obs}|\right] + \mathbb{P}\left[-T \geq |z_{obs}|\right] = \mathbb{P}\left[|T| \geq |z_{obs}|\right].$$

5.2.2 Test d'inégalité $\mu \leq \mu_0$

On veut tester au risque α l'hypothèse nulle $H0$: "$\mu \leq \mu_0$" contre l'hypothèse alternative $H1$: "$\mu > \mu_0$". Sous $H0$, il existe $\mu' \leq \mu_0$ tel que

$$Z(\mu') = \sqrt{n}\frac{\overline{X}_n - \mu'}{S_n} \sim T_{n-1}.$$

On définit la zone de rejet par

$$ZR = \left\{Z\left(\mu_0\right) > t_{n-1,1-\alpha}\right\}$$

où $t_{n-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté. On calcule alors $z_{obs} = \sqrt{n}\frac{\overline{x}_n - \mu_0}{s_n}$.
Si $z_{obs} > t_{n-1,1-\alpha}$, on rejette $H0$, sinon, on ne rejette pas $H0$.

Remarque 1 : En réalité, pour trouver notre zone de rejet, comme μ peut être aussi petit que l'on veut sous $H0$, on doit rejeter $H0$ si $\overline{X}_n$ est trop grand, i.e on cherche c_α tel que sous H0

$$\sup_{\mu \leq \mu_0} \mathbb{P}_\mu\left["\text{On rejette H0"}\right] = \sup_{\mu \leq \mu_0} \mathbb{P}_\mu\left[\overline{X}_n \geq c_\alpha\right] = \alpha$$

Or, on a

$$\begin{aligned}
\alpha &= \sup_{\mu \leq \mu_0} \mathbb{P}_\mu\left[\overline{X}_n \geq c_\alpha\right] \\
&= \sup_{\mu \leq \mu_0} \mathbb{P}_\mu\left[\overline{X}_n - \mu \geq c_\alpha - \mu\right] \\
&= \sup_{\mu \leq \mu_0} \mathbb{P}_\mu\left[\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \geq \sqrt{n}\frac{c_\alpha - \mu}{S_n}\right].
\end{aligned}$$

Comme $\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} \sim T_{n-1}$, si on note F_T la fonction de répartition de la loi de Student à $n-1$ degrés de liberté, on obtient

$$\begin{aligned}
\alpha &= \sup_{\mu \leq \mu_0}\left(1 - F_T\left(\sqrt{n}\frac{c_\alpha - \mu}{S_n}\right)\right) \\
&= 1 - \inf_{\mu \leq \mu_0} F_T\left(\sqrt{n}\frac{c_\alpha - \mu}{S_n}\right)
\end{aligned}$$

Comme la fonction F_T est strictement croissante, on obtient

$$1 - \alpha = F_T\left(\sqrt{n}\frac{c_\alpha - \mu_0}{S_n}\right)$$

et donc, en appliquant F_T^{-1},

$$t_{n-1,1-\alpha} = \sqrt{n}\frac{c_\alpha - \mu_0}{S_n}$$

et on obtient alors

$$c_\alpha = \mu_0 + t_{n-1,1-\alpha}\frac{S_n}{\sqrt{n}}.$$

Donc, sous $H0$

$$\sup_{\mu\leq\mu_0} \mathbb{P}\left[\overline{X}_n \geq c_\alpha\right] = \mathbb{P}_{\mu=\mu_0}\left[\overline{X}_n \geq \mu_0 + t_{n-1,1-\alpha}\frac{S_n}{\sqrt{n}}\right] = \mathbb{P}_{\mu=\mu_0}\left[\sqrt{n}\frac{\overline{X}_n - \mu_0}{S_n} \geq t_{n-1,1-\alpha}\right] = \alpha,$$

et on retrouve ainsi la zone de rejet.

Remarque 2 : On a

$$\begin{aligned} p-\text{value} &= \inf\{\alpha \in (0,1), \text{"On rejette H0"}\} \\ &= \inf\{\alpha \in (0,1), z_{obs} \geq t_{n-1,1-\alpha}\} \\ &= \inf\{\alpha \in (0,1), F_T(z_{obs}) \geq 1-\alpha\} \\ &= \inf\{\alpha \in (0,1), \alpha \geq 1 - F_T(z_{obs})\} \\ &= 1 - F_T(z_{obs}) \\ &= \mathbb{P}[T \geq z_{obs}], \end{aligned}$$

où $T \sim T_{n-1}$.

5.2.3 Test d'inégalité $\mu \geq \mu_0$

On veut tester au risque α l'hypothèse nulle $H0$: "$\mu \geq \mu_0$" contre l'hypothèse alternative $H1$: "$\mu < \mu_0$". Sous $H0$, il existe $\mu' \geq \mu_0$ tel que

$$Z(\mu') = \sqrt{n}\frac{\overline{X}_n - \mu'}{S_n} \sim T_{n-1}.$$

On définit la zone de rejet par

$$ZR = \{Z(\mu_0) < t_{n-1,\alpha}\}$$

où $t_{n-1,\alpha}$ est le quantile d'ordre α de la loi de Student à $n-1$ degrés de libertés. On calcule alors $z_{obs} = \sqrt{n}\frac{\overline{x}_n - \mu_0}{s_n}$. Si $z_{obs} < t_{n-1,\alpha}$, on rejette $H0$, sinon, on ne rejette pas $H0$.

Remarque 1 : En réalité, pour trouver notre zone de rejet, comme μ peut être aussi grand que l'on veut sous $H0$, on doit rejeter $H0$ si $\overline{X}_n$ est trop petit, i.e on cherche c_α tel que

$$\sup_{\mu\geq\mu_0} \mathbb{P}_\mu\left[\text{"On rejette H0"}\right] = \sup_{\mu\geq\mu_0} \mathbb{P}_\mu\left[\overline{X}_n \leq c_\alpha\right] = \alpha$$

Or, si on note F_T la fonction de répartition de la loi de Student à $n-1$ degrés de liberté, on a

$$\begin{aligned}
\alpha &= \sup_{\mu \geq \mu_0} \mathbb{P}_\mu \left[\overline{X}_n \leq c_\alpha \right] \\
&= \sup_{\mu \geq \mu_0} \mathbb{P}_\mu \left[\overline{X}_n - \mu \leq c_\alpha - \mu \right] \\
&= \sup_{\mu \geq \mu_0} \mathbb{P}_\mu \left[\sqrt{n} \frac{\overline{X}_n - \mu}{S_n} \leq \sqrt{n} \frac{c_\alpha - \mu}{S_n} \right] \\
&= \sup_{\mu \geq \mu_0} F_T \left(\sqrt{n} \frac{c_\alpha - \mu}{S_n} \right)
\end{aligned}$$

Comme la fonction F_T est strictement croissante, on obtient

$$\alpha = F_T \left(\sqrt{n} \frac{c_\alpha - \mu_0}{S_n} \right)$$

et donc, en appliquant F_T^{-1},

$$t_{n-1,\alpha} = \sqrt{n} \frac{c_\alpha - \mu_0}{S_n}.$$

Comme $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$, on obtient alors

$$c_\alpha = \mu_0 - t_{n-1,1-\alpha} \frac{S_n}{\sqrt{n}}$$

D'où, sous $H0$

$$\sup_{\mu \geq \mu_0} \mathbb{P}_\mu \left[\overline{X}_n \leq c_\alpha \right] = \mathbb{P}_{\mu=\mu_0} \left[\overline{X}_n \leq \mu_0 - t_{n-1,1-\alpha} \frac{S_n}{\sqrt{n}} \right] = \mathbb{P}_{\mu=\mu_0} \left[\sqrt{n} \frac{\overline{X}_n - \mu_0}{S_n} \leq -t_{n-1,1-\alpha} \right] = \alpha,$$

et on retrouve ainsi la zone de rejet.

Remarque 2 : On a

$$\begin{aligned}
p - \text{value} &= \inf \{ \alpha \in (0,1), \text{"On rejette H0"} \} \\
&= \inf \{ \alpha \in (0,1), z_{obs} \leq -t_{n-1,1-\alpha} \} \\
&= \inf \{ \alpha \in (0,1), F_T(z_{obs}) \leq \alpha \} \\
&= F_T(z_{obs}) \\
&= \mathbb{P}[T \leq z_{obs}],
\end{aligned}$$

où $T \sim T_{n-1}$.

5.2.4 Application :

Les données suivantes sont extraites du cours de B. Portier à l'INSA de Rouen. A la suite d'un traitement (régime alimentaire) sur une variété de porcs, on prélève un (petit) échantillon de 5

porcs et on les pèse. On obtient les poids suivants (en kg)

$$83 \quad 81 \quad 84 \quad 80 \quad 85$$

On suppose que les x_i sont des réalisations de variables aléatoires i.i.d suivant une loi normale de paramètres μ, σ^2. On sait que le poids moyen de cette variété de porcs est de $87,6$ kg et on souhaite savoir si le régime alimentaire a eu un impact ou non sur ce poids moyen, i.e on teste au risque de 5%, $H0$: "$\mu = 87,6$" contre $H1$: "$\mu \neq 87,6$". On a la statistique de test

$$Z = \sqrt{5}\frac{\overline{X}_5 - 87.6}{S_n} \sim T_4 \qquad \text{sous H0.}$$

On a donc la zone de rejet

$$ZR = \{|Z| \geq t_{4,0.975}\}.$$

Ici, $t_{4,0.975} = 2.78$. De plus, on a

$$z_{obs} = \sqrt{5}\frac{\overline{x}_5 - 87.6}{s_5} = \sqrt{5}\frac{82.6 - 87.5}{2.074} = -5.28.$$

Comme $|z_{obs}| > 2.78$, on rejette $H0$.

Un biologiste affirme que ce régime permet d'augmenter le poids des porcs, et que ces résultats ne sont dus qu'à la fluctuation d'échantillonnage. Pour le contredire, on teste alors, au risque de 5%, $H0$: "$\mu \geq 87.6$" contre $H1$: "$\mu < 87.6$". Sous $H0$, il existe $\mu' \geq 87.6$ tel que

$$\sqrt{5}\frac{\overline{X}_5 - \mu'}{S_5} \sim T_4.$$

On a la zone de rejet définie par

$$ZR = \left\{\sqrt{5}\frac{\overline{X}_5 - 87.6}{S_5} < -t_{4,0.95}\right\}$$

Ici, $t_{4,0.95} = 2.13$ et

$$\sqrt{5}\frac{\overline{x}_5 - 87.6}{s_5} = \sqrt{5}\frac{82.6 - 87.5}{2.074} = -5.28.$$

Comme $-5.28 < -2.13$, on rejette $H0$.

5.2.5 Test de conformité d'une variance

On souhaite maintenant tester, au risque α, l'hypothèse nulle $H0$: "$\sigma^2 = \sigma_0^2$" contre l'hypothèse alternative $H1$: "$\sigma^2 \neq \sigma_0^2$".

Construction de la statistique de test : On rappelle que l'on a

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Comme sous H0 $\sigma^2 = \sigma_0^2$, on a alors la statistique de test

$$\frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi_{n-1}^2 \quad \text{sous H0}$$

ce qui est bien évidemment faux sous $H1$. En effet, sous H1,

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

et ne suit donc pas une loi du Chi-deux à $n-1$ degrés de liberté.

Le test : On teste au risque α, $H0$: "$\sigma^2 = \sigma_0^2$" contre $\sigma^2 \neq \sigma_0^2$. On a la statistique de test

$$Z = \frac{n-1}{\sigma_0^2} S_n^2 \sim \chi_{n-1}^2 \qquad \text{sous H0}$$

On définit alors la zone de rejet comme

$$ZR = \{Z, Z < k_{\alpha/2}\} \cup \{Z, Z > k_{1-\alpha/2}\},$$

où $k_{\alpha/2}$ et $k_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi du chi deux à $n-1$ degrés de liberté. On calcule z_{obs}, et si z_{obs} est dans la zone de rejet, on rejette $H0$ et inversement.

Remarque 1 : Notons que faire ce test revient donc à vérifier que σ_0^2 appartient à l'intervalle de confiance obtenu à partir des données observées. Si σ_0^2 appartient à cet intervalle, alors on ne rejette pas $H0$ et inversement. En effet, on a

$$\begin{aligned}
\text{"On ne rejette pas H0"} &\Leftrightarrow k_{\alpha/2} \leq |z_{obs}| \leq k_{1-\alpha/2} \\
&\Leftrightarrow k_{\alpha/2} \leq \frac{(n-1)s_n^2}{\sigma_0^2} \leq k_{1-\alpha/2} \\
&\Leftrightarrow \frac{k_{\alpha/2}}{(n-1)s_n^2} \leq \frac{1}{\sigma_0^2} \leq \frac{k_{1-\alpha/2}}{(n-1)s_n^2} \\
&\Leftrightarrow \frac{(n-1)s_n^2}{k_{\alpha/2}} \geq \sigma_0^2 \geq \frac{(n-1)s_n^2}{k_{1-\alpha/2}} \\
&\Leftrightarrow \sigma_0^2 \in IC_{1-\alpha}\left(\sigma^2\right).
\end{aligned}$$

Remarque 2 : Soit $Z \sim \chi^2_{n-1}$, et F_Z sa fonction de répartition, alors

$$\begin{aligned} p-\text{value} &= \inf\{\alpha \in (0,1), \text{"On rejette H0"}\} \\ &= \inf\{\alpha \in (0,1), k_{\alpha/2} \geq z_{obs} \quad \text{ou} \quad z_{obs} \geq k_{1-\alpha/2}\} \\ &= \min\left(\inf\{\alpha \in (0,1), k_{\alpha/2} \geq z_{obs}\}, \inf\{\alpha \in (0,1), k_{1-\alpha/2} \leq z_{obs}\}\right) \\ &= \min\left(\inf\{\alpha \in (0,1), \alpha/2 \geq F(z_{obs})\}, \inf\{\alpha \in (0,1), 1-\alpha/2 \leq F(z_{obs})\}\right) \\ &= \min\left(2F(z_{obs}), 2-2F(z_{obs})\right) \\ &= \min\left(2\mathbb{P}[Z \leq z_{obs}], 2\mathbb{P}[Z \geq z_{obs}]\right). \end{aligned}$$

Exemple : Reprenons l'exemple des porcs. Le même biologiste pense que si les résultats ne sont pas concluants (i.e ils ne vont pas dans son sens), c'est dû à une variance qui serait, selon lui, égale à 25. Sans s'attarder sur le bien fondé de son raisonnement, on teste au risque de 5% l'hypothèse nulle H0 :"$\sigma^2 = 25$" contre l'hypothèse alternative H1 :"$\sigma^2 \neq 25$". On a la statistique de test

$$Z = \frac{4}{25} S_n^2 \sim \chi^2_4 \quad \text{sous H0}$$

On a la zone de rejet

$$ZR = \{Z, Z \leq 0.48 \text{ ou } Z \geq 11.14\}.$$

De plus, on a

$$z_{obs} = \frac{4}{25} s_n^2 = 0.29.$$

Comme $z_{obs} < 0.48$, on rejette $H0$.

5.3 Tests de comparaison de deux moyennes

5.3.1 Introduction

Présentation du problème : On considère deux jeux de données

- $x_1, \ldots, x_p$ qui sont p mesures d'une variable aléatoire X.
- $y_1, \ldots, y_q$ qui sont q mesures d'une variable aléatoire Y.

On souhaite étudier les moyennes théoriques des variables X et Y, moyennes que l'on peut respectivement estimer par $\overline{x}_p$ et $\overline{y}_q$. Bien évidemment, il est peu probable que ces deux estimations soient égales, et l'objectif est donc de savoir si cette différence est seulement due à la fluctuation d'échantillonnage, ou si elle est réellement significative.

Le modèle probabiliste : On fait les hypothèses suivantes :

- les données $x_1, \ldots, x_p$ sont les réalisations de variables aléatoires $X_1, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}(\mu_1, \sigma_1^2)$
- les données $y_1, \ldots, y_q$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}(\mu_2, \sigma_2^2)$

— les deux échantillons sont indépendants et de même variance, i.e $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Exemple : Cet exemple est issu du cours de B. Portier à l'INSA de Rouen. On souhaite comparer les productions laitières de deux races bovines. On choisit donc 50 bêtes de chaque race, i.e $p = q = 50$, et on mesure pour chaque bête la production annuelle totale de lait (en tonnes). On obtient pour ces deux échantillons les rendements moyens suivants :

$$\overline{x}_{50} = 4.860 \qquad \text{et} \qquad \overline{y}_{50} = 4.690$$

L'objectif sera donc de savoir si la différence entre ces deux moyennes est due à la fluctuation d'échantillonnage ou si il y a une réelle différence de rendement entre les deux races.

5.3.2 Test d'égalité

On veut tester au risque α l'hypothèse nulle H0 : "$\mu_1 = \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 \neq \mu_2$".

Construction de la statistique de test : On dispose des estimateurs $\overline{X}_p$ et $\overline{Y}_q$ et on s'intéresse plus particulièrement à la variable aléatoire $\overline{X}_p - \overline{Y}_q$.

Proposition 5.3.1. *On a*

$$\overline{X}_p - \overline{Y}_q \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{p} + \frac{\sigma^2}{q}\right).$$

Démonstration. On peut voir $\overline{X}_p - \overline{Y}_q$ comme une combinaison linéaire de variables aléatoires indépendantes suivant des lois normales, i.e

$$\overline{X}_p - \overline{Y}_q = \frac{1}{p}\sum_{i=1}^{p} X_i - \frac{1}{q}\sum_{i=1}^{q} Y_i,$$

et suit donc une loi normale avec

$$\mathbb{E}\left[\overline{X}_p - \overline{Y}_q\right] = \frac{1}{p}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] - \frac{1}{q}\sum_{i=1}^{n} \mathbb{E}\left[Y_i\right] = \mu_1 - \mu_2$$

et comme $\overline{X}_n$ et $\overline{Y}_n$ sont indépendantes,

$$\mathbb{V}\left[\overline{X}_p - \overline{Y}_q\right] = \mathbb{V}\left[\overline{X}_p\right] + \mathbb{V}\left[\overline{Y}_q\right] = \frac{\sigma^2}{p} + \frac{\sigma^2}{q}.$$

□

On obtient ainsi

$$\frac{\overline{X}_p - \mu_1 + \mu_2 - \overline{Y}_q}{\sigma\sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{X}_p - \mu_1 + \mu_2 - \overline{Y}_q}{\sigma} \sim \mathcal{N}(0,1).$$

Malheureusement, on ne connait pas nécessairement σ^2 et donc pas tous les paramètres de la loi de $\overline{X}_p - \overline{Y}_q$. Cependant, on peut le remplacer par un estimateur. La proposition suivante donne cet estimateur ainsi que sa loi.

Proposition 5.3.2. *On considère S^2 l'estimateur de σ^2 défini par*

$$S^2 = \frac{1}{p+q-2}\left(\sum_{i=1}^{p}\left(X_i - \overline{X}_p\right)^2 + \sum_{i=1}^{q}\left(Y_i - \overline{Y}_q\right)^2\right) = \frac{1}{p+q-2}\left((p-1)S_X^2 + (q-1)S_Y^2\right),$$

avec $S_X^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(X_i - \overline{X}_p\right)^2$ et $S_Y^2 = \frac{1}{q-1}\sum_{i=1}^{q}\left(Y_i - \overline{Y}_p\right)^2$.

1. *On a*
$$\frac{(p+q-2)S^2}{\sigma^2} \sim \chi^2_{p+q-2}.$$
2. *Les variables $\overline{X}_p - \overline{Y}_q$ et S^2 sont indépendantes.*

Démonstration. Comme

$$\frac{(p-1)S_X^2}{\sigma^2} \sim \chi^2_{p-1} \qquad \text{et} \qquad \frac{(q-1)S_Y^2}{\sigma^2} \sim \chi^2_{q-1},$$

en appliquant la proposition 4.3.1, on obtient

$$\frac{p+q-2}{\sigma^2}S^2 = \frac{(p-1)S_X^2}{\sigma^2} + \frac{(q-1)S_Y^2}{\sigma^2} \sim \chi^2_{p+q-2}.$$

De plus comme $S_X^2, \overline{X}_q, S_Y^2$ et $\overline{Y}_q$ sont indépendants, S^2 et $\overline{X}_p - \overline{Y}_q$ sont indépendants. □

Version 2. On note

$$Z = \begin{pmatrix} X_1 - \mu_1 \\ \vdots \\ X_p - \mu_1 \\ Y_1 - \mu_2 \\ \vdots \\ Y_q - \mu_2 \end{pmatrix} \sim \mathcal{N}\left(0, \sigma^2 I_{p+q}\right).$$

L'objectif est donc d'utiliser le théorème de Cochran et donc de trouver une décomposition en sous-espaces orthogonaux adéquate. On note $\{e_1, \ldots, e_{p+q}\}$ la base canonique de $\mathbb{R}^{p+q}$, et on considère les sous espaces vectoriels

$$V = \text{vect}\{e_1, \ldots, e_p\} \qquad \text{et} \qquad W = \text{vect}\{e_{p+1}, \ldots, e_{p+q}\}.$$

A noter que l'on a $\mathbb{R}^{p+q} = V \oplus_\perp W$ et que les projecteurs orthogonaux sur les sous-espaces sont

définis pour tout h par

$$P_V : h \longmapsto \sum_{i=1}^{p} \langle e_i, h \rangle e_i \qquad \text{et} \qquad P_W : h \longmapsto \sum_{i=p+1}^{p+q} \langle e_i, h \rangle e_i.$$

On considère également $v = \frac{1}{\sqrt{p}} \left(\mathbf{1}_p^T, 0_q^T \right)^T$ avec $\mathbf{1}_p^T = \underbrace{(1, \ldots, 1)}_{p}$ et $0_q^T = \underbrace{(0, \ldots, 0)}_{q}$, ainsi que $w = \frac{1}{\sqrt{q}} \left(0_p, \mathbf{1}_q \right)^T$. On note $V_1 = \text{vect}\{v\}$ et $W_1 = \text{vect}\{w\}$ et on a donc V_1 (resp. W_1) qui est un sous-espace vectoriel de V (resp. W) et on note $V_{1,\perp}$ (resp. $W_{2,\perp}$) le sous-espace vectoriel tel que $V = V_1 \oplus_\perp V_{1,\perp}$ (resp. $W = W_1 \oplus_\perp W_{1,\perp}$). A noter que les projections orthogonales sont donc données par

$$P_{V_1} = vv^T : h \longmapsto \left(v^T h \right) v \qquad P_{W_1} = ww^T : h \longmapsto \left(w^T h \right) w$$

On obtient donc la décomposition suivante

$$\mathbb{R}^{p+q} = V_1 \oplus_\perp V_{1,\perp} \oplus_\perp W_1 \oplus_\perp W_{1,\perp}.$$

Il ne reste donc plus qu'à calculer les projections. On a facilement

$$P_{V_1}(Z) = \left(\overline{X}_p - \mu_1 \right) (1, \ldots, 1, 0, \ldots, 0)^T \qquad \text{et} \qquad P_{W_1} = \left(\overline{Y}_q - \mu_q \right) (0, \ldots, 0, 1, \ldots, 1)^T.$$

De plus, on a

$$\begin{aligned} P_{V_{1,\perp}}(Z) = P_V(Z) - P_{V_1}(Z) &= \left((X_1 - \mu_1), \ldots, \left(X_p - \mu_1 \right), 0, \ldots, 0 \right)^T - \left(\overline{X}_p - \mu_1 \right) (1, \ldots, 1, 0, \ldots, 0)^T \\ &= \left(\left(X_1 - \overline{X}_p \right), \ldots, \left(X_p - \overline{X}_p \right), 0, \ldots, 0 \right)^T \end{aligned}$$

De la même façon, on a

$$P_{W'_{1,\perp}(Z)} = \left(0, \ldots, 0, Y_1 - \overline{Y}_q, \ldots, Y_q - \overline{Y}_q \right)^T.$$

Ainsi, en appliquant le théorème de Cochran,

— $\overline{X}_n, \overline{Y}_n, P_{V_{1,\perp}}(Z), P_{W_{1,\perp}}(Z)$ sont indépendants.
— On a

$$\frac{1}{\sigma^2} \left\| P_{V_{1,\perp}} \right\|^2 = \frac{1}{\sigma^2} \sum_{i=1}^{p} \left(X_i - \overline{X}_p \right)^2 = \frac{p-1}{\sigma^2} S_X^2 \sim \chi^2_{p-1}$$
$$\frac{1}{\sigma^2} \left\| P_{W_{1,\perp}} \right\|^2 = \frac{1}{\sigma^2} \sum_{i=1}^{q} \left(Y_i - \overline{Y}_q \right)^2 = \frac{q-1}{\sigma^2} S_Y^2 \sim \chi^2_{q-1}$$

On a donc indépendance de $S^2, \overline{X}_p$ et $\overline{Y}_q$. Comme $P_{V_{1,\perp}}(Z)$ et $P_{W_{1,\perp}}(Z)$ sont indépendants,

en appliquant la proposition 4.3.1, on obtient

$$\frac{1}{\sigma^2}\left((p-1)S_X^2+(q-1)S_Y^2\right)\sim\chi^2_{p+q-2}$$

□

Version 3. On considère U le sous-espace vectoriel tel que

$$\mathbb{R}^{p+q}=V_1\oplus_\perp W_1\oplus_\perp U.$$

La projection orthogonale de Z sur U est donnée par

$$P_U(Z)=Z-P_{V_1}(Z)-P_{W_1}(Z)=\left(\left(X_1-\overline{X}_p\right),\ldots,\left(X_p-\overline{X}_p\right),\left(Y_1-\overline{Y}_q\right),\ldots,\left(Y_q-\overline{Y}_q\right)\right).$$

Ainsi

$$\|P_U(Z)\|^2=\sum_{i=1}^{p}\left(X_i-\overline{X}_p\right)^2+\sum_{i=1}^{q}\left(Y_i-\overline{Y}_q\right)^2=(p+q-2)S^2.$$

En appliquant le théorème de Cochran, on obtient donc

— $\overline{X}_p,\overline{Y}_q$ et S^2 sont indépendants.
— $\frac{1}{\sigma^2}\left(\sum_{i=1}^{p}(X_i-\mu_1)^2+\sum_{i=1}^{q}(Y_i-\mu_2)^2\right)\sim\chi^2_{p+q-2}$.

□

Le corollaire suivant est une application directe de la proposition précédente.

Corollaire 5.3.1. *On a*

$$\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\left(\overline{X}_p-\overline{Y}_q\right)-(\mu_1-\mu_2)}{S}\sim T_{p+q-2}$$

où T_{p+q-2} est une loi de Student à $p+q-2$ degrés de liberté.

Démonstration. On a

$$\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\left(\overline{X}_p-\overline{Y}_q\right)-(\mu_1-\mu_2)}{S}=\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\left(\overline{X}_p-\overline{Y}_q\right)-(\mu_1-\mu_2)}{\sigma}\frac{1}{\sqrt{\frac{(p+q-2)S^2}{\sigma^2(p+q-2)}}}\sim T_{p+q-2}$$

car S^2 et $\overline{X}_p-\overline{Y}_q$ sont indépendants. □

On obtient donc la statistique de test

$$\frac{\sqrt{pq}\left(\overline{X}_p-\overline{Y}_q\right)}{S\sqrt{p+q}}\sim T_{p+q-2}\quad\text{sous H0,}$$

ce qui est évidemment faux sous H1 car alors $\mu_1-\mu_2\neq 0$.

Le test :

— On teste au risque α

$$H0 : "\mu_1 = \mu_2" \qquad \text{contre} \qquad H1 : "\mu_1 \neq \mu_2".$$

— On a la statistique de test

$$Z = \frac{\sqrt{pq}\left(\overline{X}_p - \overline{Y}_q\right)}{S\sqrt{p+q}} \sim T_{p+q-2} \quad \text{sous H0.}$$

— On a alors la zone de rejet :

$$ZR = \left\{Z, |Z| > t_{p+q-2,1-\alpha/2}\right\}$$

où $t_{p+q-2,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $p+q-2$ degrés de liberté.

— Si $|z_{obs}| > t_{p+q-2,1-\alpha/2}$, on rejette H0 et inversement.

Remarque 1 : Cela revient à vérifier que 0 appartient à l'intervalle de confiance de niveau $1-\alpha$ de $\mu = \mu_1 - \mu_2$. En effet

$$\begin{aligned}
|z_{obs}| \leq t_{p+q-2,1-\alpha/2} &\Leftrightarrow -t_{p+q-2,1-\alpha/2} \leq \frac{\sqrt{pq}\left(\overline{x}_p - \overline{y}_q\right)}{s\sqrt{p+q}} \leq t_{p+q-2,1-\alpha/2} \\
&\Leftrightarrow \overline{x}_p - \overline{y}_q - \frac{st_{p+q-2,1-\alpha/2}\sqrt{p+q}}{\sqrt{pq}} \leq 0 \leq \overline{x}_p - \overline{y}_q + \frac{st_{p+q-2,1-\alpha/2}\sqrt{p+q}}{\sqrt{pq}} \\
&\Leftrightarrow 0 \in IC_{1-\alpha}\left(\mu_1 - \mu_2\right).
\end{aligned}$$

Remarque 2 : En réalité, pour trouver la zone de rejet, comme $\overline{X}_p - \overline{Y}_q$ doit être proche de 0, on cherche c_α tel que

$$\mathbb{P}_{\mu_1=\mu_2}\left["\text{On rejette H0}"\right] = \mathbb{P}_{\mu_1=\mu_2}\left[\left|\overline{X}_p - \overline{Y}_q\right| \geq c_\alpha\right] = \alpha.$$

On obtient donc

$$\begin{aligned}
\alpha &= \mathbb{P}_{\mu_1=\mu_2}\left[\left|\overline{X}_p - \overline{Y}_q\right| \geq c_\alpha\right] \\
&= \mathbb{P}_{\mu_1=\mu_2}\left[\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\left|\overline{X}_p - \overline{Y}_q\right|}{S} \geq \frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha}{S}\right],
\end{aligned}$$

et on obtient donc, en notant F_T la fonction de répartition de la loi de Student à $p+q-2$ degrés de liberté, sous H0,

$$\alpha = 2 - 2F_T\left(\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha}{S}\right).$$

En appliquant F_T^{-1}, on obtient

$$t_{p+q-2,1-\alpha/2} = \frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha}{S},$$

et donc

$$c_\alpha = t_{p+q-2,1-\alpha/2}\frac{\sqrt{p+q}}{\sqrt{pq}}S.$$

On retrouve alors la zone de rejet, i.e

$$ZR = \left\{ \left|\overline{X}_p - \overline{Y}_q\right| \geq t_{p+q-2,1-\alpha/2}\frac{\sqrt{p+q}}{\sqrt{pq}}S \right\} = \left\{ \frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\left|\overline{X}_p - \overline{Y}_q\right|}{S} \geq t_{p+q-2,1-\alpha/2} \right\}.$$

Remarque 3 : Soit $T \sim T_{p+q-2}$, on a alors

$$p - \text{value} = \mathbb{P}\left[|T| > |z_{obs}|\right].$$

En effet, on note F_T la fonction de répartition de T,

$$\begin{aligned} p - \text{value} &= \inf\{\alpha \in (0,1), \text{"On rejette H0"}\} \\ &= \inf\left\{\alpha \in (0,1), |z_{obs}| \geq t_{p+q-2,1-\alpha/2}\right\} \\ &= \inf\left\{\alpha \in (0,1), F_T\left(|z_{obs}|\right) \geq 1-\alpha/2\right\} \\ &= 2 - 2F_T\left(|z_{obs}|\right) \\ &= \mathbb{P}\left[|T| \geq |z_{obs}|\right), \end{aligned}$$

la dernière égalité résultant de la symétrie de la loi de Student.

Exemple : les Iris de Fisher On considère un jeu de données "classique". L'objectif est de reconnaître le type d'Iris à partir de la longueur de ses sépales. On considère ici deux espèces, et pour chaque espèce on dispose de 50 individus. Plus précisément, on note $x_1, \ldots, x_{50}$ les longueurs des sépales des iris de la variété Virginica et $y_1, ..., y_{50}$ celle de la variété Versicolor. On obtient les résultat suivants

$$\begin{aligned} \overline{x}_{50} &= 5.94 & s_X &= 0.52 \\ \overline{y}_{50} &= 6.59 & s_Y &= 0.64 \end{aligned}$$

On teste au risque 5% l'hypothèse nulle H0 : "$\mu_1 = \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 \neq \mu_2$". On a la statistique de test

$$Z = \frac{\sqrt{50 \times 50}\left(\overline{X}_{50} - \overline{Y}_{50}\right)}{S\sqrt{50+50}} \sim T_{98} \quad \text{sous H0.}$$

Ici $z_{obs} = -5.63$. On a la zone de rejet

$$ZR = \left\{|Z| > t_{98,0.975}\right\}.$$

Ici $t_{98,0.975} = 1.98$. Donc $|z_{obs}| > t_{98,0.975}$ et on rejette donc H0, i.e on peut donc penser que la longueur moyenne des sépales des Iris dépend de la variété.

5.3.3 Test d'inégalité $\mu_1 \leq \mu_2$

On veut tester au risque α l'hypothèse nulle H0 : "$\mu_1 \leq \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 > \mu_2$". Sous H0, il existe $\mu' \leq 0$ tel que

$$Z(\mu') = \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{X}_p - \overline{Y}_q - \mu'}{S} \sim T_{p+q-2}.$$

On définit alors la zone de rejet par

$$ZR = \left\{ Z(0) > t_{p+q-2,1-\alpha} \right\},$$

où $t_{p+q-2,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $p+q-2$ degrés de liberté. On calcule $z_{obs} = \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{x}_p - \overline{y}_q}{s}$. Si $z_{obs} > t_{p+q-2,1-\alpha}$, on rejette H0 et inversement.

Remarque 1 : En réalité, pour trouver la zone de rejet, comme $\mu = \mu_1 - \mu_2$ peut être aussi petit que l'on veut sous H0, on doit rejeter H0 si $\overline{X}_p - \overline{Y}_q$ est trop grand, i.e on cherche c_α tel que

$$\sup_{\mu \leq 0} \mathbb{P}_\mu \left["\text{On rejette H0"}\right] = \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\overline{X}_p - \overline{Y}_q \geq c_\alpha\right] = \alpha.$$

Or si on note F_T la fonction de répartition de la loi de Student à $p+q-2$ degrés de liberté, on a

$$\begin{aligned}
\alpha &= \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\overline{X}_p - \overline{Y}_q \geq c_\alpha\right] \\
&= \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\overline{X}_p - \overline{Y}_q - \mu \geq c_\alpha - \mu\right] \\
&= \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{X}_p - \overline{Y}_q - \mu}{S} \geq \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{c_\alpha - \mu}{S}\right] \\
&= \sup_{\mu \leq 0} \left(1 - F_T\left(\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{c_\alpha - \mu}{S}\right)\right)
\end{aligned}$$

Comme la fonction F_T est strictement croissante, on obtient

$$\alpha = 1 - F_T\left(\frac{\sqrt{pq}}{\sqrt{p+q}} \frac{c_\alpha}{S}\right)$$

et donc, en appliquant F_T^{-1},

$$t_{p+q-2,1-\alpha} = \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{c_\alpha}{S},$$

et donc

$$c_\alpha = \frac{\sqrt{p+q}}{\sqrt{pq}} t_{p+q-2,1-\alpha} S$$

On retrouve ainsi la zone de rejet

$$ZR = \left\{ \overline{X}_p - \overline{Y}_q \geq t_{p+q-2,1-\alpha} \frac{\sqrt{p+q}}{\sqrt{pq}} S \right\} = \left\{ \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{X}_p - \overline{Y}_q}{S} \geq t_{p+q-2,1-\alpha} \right\}.$$

Remarque 2 : On a

$$p - \text{value} = \mathbb{P}\left[T \geq z_{obs}\right],$$

où $T \sim T_{p+q-2}$. En effet,

$$\begin{aligned} p - \text{value} &= \inf\left\{\alpha \in (0,1), \text{"On rejette H0"}\right\} \\ &= \inf\left\{\alpha \in (0,1), z_{obs} \geq t_{p+q-2,1-\alpha}\right\} \\ &= \inf\left\{\alpha \in (0,1), F_T\left(z_{obs}\right) \geq 1-\alpha\right\} \\ &= 1 - F_T\left(z_{obs}\right). \end{aligned}$$

5.3.4 Test d'inégalité $\mu_1 \geq \mu_2$

On veut tester au risque α l'hypothèse nulle H0 : "$\mu_1 \geq \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 < \mu_2$". Sous H0, il existe $\mu' \geq 0$ tel que

$$Z\left(\mu'\right) = \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{X}_p - \overline{Y}_q - \mu'}{S} \sim T_{p+q-2}.$$

On définit alors la zone de rejet par

$$ZR = \left\{Z\left(0\right) < -t_{p+q-2,1-\alpha}\right\},$$

où $t_{p+q-2,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $p+q-2$ degrés de liberté. On calcule $z_{obs} = \frac{\sqrt{pq}}{\sqrt{p+q}} \frac{\overline{x}_p - \overline{y}_q}{s}$. Si $z_{obs} < -t_{p+q-2,1-\alpha}$, on rejette H0 et inversement.

Remarque 1 : En réalité, pour trouver la zone de rejet, comme $\mu = \mu_1 - \mu_2$ peut être aussi grand que l'on veut sous H0, on doit rejeter H0 si $\overline{X}_p - \overline{Y}_q$ est trop petit, i.e on cherche c_α tel que

$$\sup_{\mu \geq 0} \mathbb{P}_\mu\left[\text{"On rejette H0"}\right] = \sup_{\mu \geq 0} \mathbb{P}_\mu\left[\overline{X}_p - \overline{Y}_q \leq c_\alpha\right] = \alpha.$$

Or si on note F_T la fonction de répartition de la loi de Student à $p+q-2$ degrés de liberté, on a

$$\begin{aligned}
\alpha &= \sup_{\mu\geq 0} \mathbb{P}_\mu \left[\overline{X}_p - \overline{Y}_q \leq c_\alpha\right] \\
&= \sup_{\mu\geq 0} \mathbb{P}_\mu \left[\overline{X}_p - \overline{Y}_q - \mu \leq c_\alpha - \mu\right] \\
&= \sup_{\mu\geq 0} \mathbb{P}_\mu \left[\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\overline{X}_p - \overline{Y}_q - \mu}{S} \leq \frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha - \mu}{S}\right] \\
&= \sup_{\mu\geq 0} F_T\left(\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha - \mu}{S}\right)
\end{aligned}$$

Comme la fonction F_T est strictement croissante, on obtient

$$\alpha = F_T\left(\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha}{S}\right)$$

et donc, en appliquant F_T^{-1},

$$t_{p+q-2,\alpha} = \frac{\sqrt{pq}}{\sqrt{p+q}}\frac{c_\alpha}{S},$$

et donc, comme $t_{p+q-2,\alpha} = -t_{p+q-2,1-\alpha}$,

$$c_\alpha = -\frac{\sqrt{p+q}}{\sqrt{pq}} t_{p+q-2,1-\alpha} S.$$

On retrouve ainsi la zone de rejet,

$$ZR = \left\{\overline{X}_p - \overline{Y}_q \leq -t_{p+q-2,1-\alpha}\frac{\sqrt{p+q}}{\sqrt{pq}}S\right\} = \left\{\frac{\sqrt{pq}}{\sqrt{p+q}}\frac{\overline{X}_p - \overline{Y}_q}{S} \leq t_{p+q-2,\alpha}\right\}.$$

Remarque 2 : On a

$$p - \text{value} = \mathbb{P}\left[T \leq z_{obs}\right],$$

où $T \sim T_{p+q-2}$. En effet,

$$\begin{aligned}
p - \text{value} &= \inf\left\{\alpha \in (0,1), \text{"On rejette H0"}\right\} \\
&= \inf\left\{\alpha \in (0,1), z_{obs} \leq t_{p+q-2,\alpha}\right\} \\
&= \inf\left\{\alpha \in (0,1), F_T\left(z_{obs}\right) \leq \alpha\right\} \\
&= F_T\left(z_{obs}\right).
\end{aligned}$$

5.4 Test de Fischer et test de Shapiro-Wilk

On rappelle que pour faire les tests de Student, on a du faire les hypothèses suivantes :

— les données $x_1, \ldots, x_p$ sont les réalisations de variables aléatoires $X_1, \ldots, X_p$ indépendantes

et de même loi $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$
— les données $y_1, \ldots, y_q$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$
— les deux échantillons sont indépendants et de même variance, i.e $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

L'indépendance des variables est souvent dépendante du protocole expérimental. L'objectif de cette section et des prochaines est de donner des tests permettant de vérifier chacune des autres hypothèses, i.e que les deux variables X et Y suivent des lois normales, et qu'elles ont la même variance.

5.4.1 Test de Fischer

On veut tester au risque α l'hypothèse nulle H0 : "$\sigma_1^2 = \sigma_2^2$" contre l'hypothèse alternative H1 : "$\sigma_1^2 \neq \sigma_2^2$.

Le cadre probabiliste :
— les données $x_1, \ldots, x_p$ sont les réalisations de variables aléatoires $X_1, \ldots, X_p$ indépendantes et de même loi $\mathcal{N}\left(\mu_1, \sigma_1^2\right)$
— les données $y_1, \ldots, y_q$ sont les réalisations de variables aléatoires $Y_1, \ldots, Y_q$ indépendantes et de même loi $\mathcal{N}\left(\mu_2, \sigma_2^2\right)$
— les deux échantillons sont indépendants

Construction de la statistique de test : On dispose des estimateurs de σ_1^2 et σ_2^2 définis par

$$S_X^2 = \frac{1}{p-1}\sum_{i=1}^{p}\left(X_i - \overline{X}_p\right)^2 \quad \text{et} \quad S_Y^2 = \frac{1}{q-1}\sum_{i=1}^{q}\left(Y_i - \overline{Y}_q\right)^2.$$

La construction du test de Fisher repose sur la définition suivante.

Definition 5.4.1. *Soient p, q deux entiers positifs et soit $Z_p \sim \chi_p^2$ et $Z_q \sim \chi_q^2$ deux variables aléatoires indépendantes. Alors*

$$\frac{\frac{Z_p}{p}}{\frac{Z_q}{q}} \sim F\left(p, q\right)$$

où $F(p, q)$ suit une loi de Fisher à p, q degrés de liberté.

Corollaire 5.4.1. *On a*

$$\frac{\sigma_2^2 S_X^2}{\sigma_1^2 S_Y^2} \sim F(p-1, q-1).$$

Démonstration. Comme les variables aléatoires $X_1, \ldots, X_p$ et $Y_1, \ldots, Y_q$ sont indépendantes, les variables S_X^2 et S_Y^2 sont indépendantes. De plus

$$\frac{p-1}{\sigma_1^2} S_X^2 \sim \chi_{p-1}^2 \qquad \frac{q-1}{\sigma_2^2} S_Y^2 \sim \chi_{q-1}^2,$$

ce qui conclue la preuve. □

Comme sous H0, $\sigma_1^2 = \sigma_2^2$, on a

$$\frac{S_X^2}{S_Y^2} \sim F(p-1, q-1) \quad \text{sous H0,}$$

ce qui est bien évidemment faux sous H1.

Le test :

— On teste au risque α l'hypothèse nulle H0 : "$\sigma_1^2 = \sigma_2^2$" contre H1 : "$\sigma_1^2 \neq \sigma_2^2$".

— On a la statistique de test

$$Z = \frac{S_X^2}{S_Y^2} \sim F(p-1, q-1) \quad \text{sous H0.}$$

— On a la zone de rejet

$$ZR = \left\{Z < f_{p,q,\alpha/2}\right\} \cup \left\{Z > f_{p,q,1-\alpha/2}\right\},$$

où $f_{p,q,\alpha/2}$ et $f_{p,q,1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi de Fisher à $p-1, q-1$ degrés de liberté.

— On calcule z_{obs}. Si z_{obs} appartient à la zone de rejet, on rejette H0 et inversement.

Remarque 1 : A noter que pour tout $\alpha \in (0,1)$ et pour tout entiers positifs p, q, on a

$$f_{p,q,\alpha} = \frac{1}{f_{p,q,1-\alpha}}.$$

Remarque 2 : On peut intervertir S_X^2 et S_Y^2 et on a alors la statistique de test

$$Z = \frac{S_Y^2}{S_X^2} \sim F\left(q-1, p-1\right) \quad \text{sous H0.}$$

Remarque 3 : Soit F une variable aléatoire suivant une loi de Fisher de paramètres p, q, on a

$$p - \text{value} = \min\left\{\mathbb{P}\left[F \leq z_{obs}\right], \mathbb{P}\left[F \geq z_{obs}\right]\right\}.$$

En effet, notons $F_{p,q}$ la fonction de répartition de la loi de Fisher qui est strictement croissante, on a

$$\begin{aligned}
p - \text{value} &= \inf\left\{\alpha \in (0,1), z_{obs} < f_{p,q,\alpha} \quad \text{ou} \quad z_{obs} > f_{p,q,1-\alpha}\right\} \\
&= \min\left\{\inf\left\{\alpha \in (0,1), z_{obs} < f_{p,q,\alpha}\right\}, \inf\left\{\alpha \in (0,1), z_{obs} > f_{p,q,1-\alpha}\right\}\right\} \\
&= \min\left\{\inf\left\{\alpha \in (0,1), F_{p,q}\left(z_{obs}\right) < \alpha\right\}, \inf\left\{\alpha \in (0,1), F_{p,q}\left(z_{obs}\right) > 1-\alpha\right\}\right\} \\
&= \min\left\{F_{p,q}\left(z_{obs}\right), 1 - F_{p,q}\left(z_{obs}\right)\right\}
\end{aligned}$$

Remarque 4 : On peut rencontrer la statistique de test

$$Z = \frac{\max\left(S_X^2, S_Y^2\right)}{\min\left(S_X^2, S_Y^2\right)}.$$

Exemple : On reprend l'exemple des iris de Fisher. Rappelons que l'on a $s_X^2 = 0.52^2$ et $s_Y^2 = 0.64^2$. On teste au risque de 5% l'hypothèse nulle H0 : "$\sigma_1^2 = \sigma_2^2$" contre l'hypothèse alternative H1 : "$\sigma_1^2 \neq \sigma_2^2$". On a la statistique de test

$$Z = \frac{S_X^2}{S_Y^2} \sim F(49, 49) \quad \text{sous H0}$$

On a la zone de rejet

$$Z = \{Z < 0.57\} \cup \{Z > 1.76\}.$$

Ici, $z_{obs} = 0.66$ et on ne rejette donc pas $H0$, i.e on ne rejette pas le fait que les variances puissent être égales.

5.4.2 Test de Shapiro-Wilk

Afin de continuer de vérifier le cadre probabiliste pour les tests de Fisher, il reste entre autre à vérifier que les variables aléatoires X et Y suivent des lois normales. On considère $x_1, \ldots, x_n$ qui sont des réalisations des variables aléatoires indépendantes et de même loi $X_1, \ldots, X_n$. Le test de Shapiro-Wilk permet de tester l'hypothèse nulle H0 : "La variable X est gaussienne" contre l'hypothèse alternative H1 : "La variable X n'est pas gaussienne". On ne donnera pas ici les détails de ce test, on retiendra seulement la conclusion :

- si la p-value est supérieure à α, alors on ne peut pas rejeter H0, i.e on ne peut pas rejeter le fait que les données soient distribuées selon une loi normale.
- sinon, on rejette H0.

Exemple : Si on reprend l'exemple des iris, on obtient (pour les espèces Versicolor et Virginica) pour le test de Shapiro-Wilk les p-values suivantes : 0.46 et 0.26. On ne rejette donc pas $H0$, i.e on ne rejette pas le caractère gaussien des données observées.

5.5 Test de Student dans le cas apparié

5.5.1 Introduction

On considère des données $((x_1, y_1), \ldots, (x_n, y_n))$ qui sont des réalisations de couples de variables aléatoires indépendants $(X_1, Y_1), \ldots, (X_n, Y_n)$ de même loi que (X, Y). On souhaite comparer les moyennes des variables aléatoires X et Y.

Le cadre probabiliste

— Les variables aléatoires X_i et Y_i ne sont pas (nécessairement) indépendantes.
— Les espérances de X et Y sont données par μ_1 et μ_2.
— La variable aléatoire $X - Y$ suit une loi normale d'espérance $\mu = \mu_1 - \mu_2$ et de variance σ^2.

5.5.2 Test d'égalité

On veut tester au risque α l'hypothèse nulle H0 : "$\mu_1 = \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 \neq \mu_2$".

Construction de la statistique de test : On dispose des estimateurs $\overline{X}_n$ et $\overline{Y}_n$ et on s'intéresse à la variable aléatoire $\overline{X}_n - \overline{Y}_n$.

Proposition 5.5.1. *On a*

$$\overline{X}_n - \overline{Y}_n \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma^2}{n}\right)$$

Démonstration. On a $\overline{X}_n - \overline{Y}_n = \frac{1}{n}\sum_{i=1}^n (X_i - Y_i)$ qui est une combinaison linéaire de variables aléatoires indépendantes suivant des lois normales et par linéarité

$$\mathbb{E}\left[\overline{X}_n - \overline{Y}_n\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_i\right] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Y_i\right] = \mu_1 - \mu_2,$$

et par indépendance

$$\mathbb{V}\left[\overline{X}_n - \overline{Y}_n\right] = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}\left[X_i - Y_i\right] = \frac{\sigma^2}{n}.$$

□

On obtient ainsi

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - (\mu_1 - \mu_2)}{\sigma} \sim \mathcal{N}(0,1).$$

Cependant, on ne connait pas la variance σ^2, que l'on va donc remplacer par un estimateur. La proposition suivante nous donne cet estimateur ainsi que sa loi.

Proposition 5.5.2. *On considère l'estimateur S^2 de σ^2 défini par*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - Y_i - \left(\overline{X}_n - \overline{Y}_n\right)\right)^2.$$

Cet estimateur vérifie alors

1. *On a*
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$
2. *Les variables $\overline{X}_n - \overline{Y}_n$ et S^2 sont indépendantes.*

Démonstration. Il suffit juste de considérer la variable aléatoire $Z = X - Y$, $\mu = \mu_1 - \mu_2$. En remarquant que

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Z_i - \overline{Z}_n\right)^2,$$

on obtient

1. $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
2. $\frac{(n-1)S^2}{\sigma^2}$ et $\overline{Z}_n$ sont indépendants.

□

Corollaire 5.5.1. *On a*

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - (\mu_1 - \mu_2)}{S} \sim T_{n-1}$$

où T_{n-1} suit une loi de Student à $n-1$ degrés de liberté.

On obtient donc la statistique de test

$$\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n}{S} \sim T_{n-1} \quad \text{sous H0,}$$

ce qui est bien évidemment faux sous H1.

Le test :

— On teste au risque α l'hypothèse nulle H0 : "$\mu_1 = \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 \neq \mu_2$".

— On a la statistique de test

$$Z = \sqrt{n}\frac{\overline{X}_n - \overline{Y}_n}{S} \sim T_{n-1} \quad \text{sous H0.}$$

— On a la zone de rejet

$$ZR = \{Z, |Z| > t_{n-1,1-\alpha/2}\}$$

où $t_{n-1,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-1$ degrés de liberté.

— Si $|z_{obs}| > t_{n-1,1-\alpha/2}$, on rejette H0 et inversement.

Remarque 1 : Cela revient à vérifier que 0 appartient à l'intervalle de confiance de niveau $1-\alpha$ de $\mu = \mu_1 - \mu_2$. En effet

$$\begin{aligned}
|z_{obs}| \leq t_{n-1,1-\alpha/2} &\Leftrightarrow -t_{n-1,1-\alpha/2} \leq \sqrt{n}\frac{\overline{x}_n - \overline{y}_n}{s} \leq t_{n-1,1-\alpha/2} \\
&\Leftrightarrow \overline{x}_n - \overline{y}_n - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq 0 \leq \overline{x}_n - \overline{y}_n + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \\
&\Leftrightarrow \overline{z}_n - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \leq 0 \leq \overline{z}_n + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}} \\
&\Leftrightarrow 0 \in IC_{1-\alpha}(\mu).
\end{aligned}$$

Remarque 2 : En réalité, pour trouver la zone de rejet, comme $\overline{X}_n - \overline{Y}_n$ doit être proche de 0, on cherche c_α tel que

$$\mathbb{P}_{\mu=0}\left["\text{On rejette H0}"\right] = \mathbb{P}_{\mu=0}\left[\left|\overline{X}_n - \overline{Y}_n\right| \geq c_\alpha\right] = \alpha.$$

On obtient donc, par symétrie de la loi de Student

$$\begin{aligned}
\alpha &= \mathbb{P}_{\mu=0}\left[\left|\overline{X}_n - \overline{Y}_n\right| \geq c_\alpha\right] \\
&= \mathbb{P}_{\mu=0}\left[\sqrt{n}\frac{\left|\overline{X}_n - \overline{Y}_n\right|}{S} \geq \frac{\sqrt{n}}{S}c_\alpha\right] \\
&= \mathbb{P}_{\mu=0}\left[\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n}{S} \geq \frac{\sqrt{n}}{S}c_\alpha\right] + \mathbb{P}_{\mu=0}\left[\sqrt{n}\frac{\overline{X}_n - \overline{Y}_q}{S} \leq -\frac{\sqrt{n}}{S}c_\alpha\right] \\
&= 2\mathbb{P}_{\mu=0}\left[\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n}{S} \geq \frac{\sqrt{n}}{S}c_\alpha\right].
\end{aligned}$$

On obtient donc, en notant F_T la fonction de répartition de la loi de Student à $n-1$ degrés de liberté,

$$\alpha = 2 - 2F_T\left(\frac{\sqrt{n}}{S}c_\alpha\right).$$

En appliquant F_T^{-1}, on obtient

$$t_{n-1,1-\alpha/2} = \frac{\sqrt{n}}{S}c_\alpha$$

et donc

$$c_\alpha = t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}.$$

On retrouve ainsi la zone de rejet

$$ZR = \left\{\left|\overline{X}_n - \overline{Y}_n\right| \geq t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right\} = \left\{\sqrt{n}\frac{\left|\overline{X}_n - \overline{Y}_n\right|}{S} \geq t_{n-1,1-\alpha/2}\right\}.$$

Remarque 3 : Soit $T \sim T_{n-1}$, on a alors

$$\begin{aligned}
p - \text{value} &= \inf\left\{\alpha \in (0,1), "\text{On rejette H0}"\right\} \\
&= \inf\left\{\alpha \in (0,1), |z_{obs}| \geq t_{n-1,1-\alpha/2}\right\} \\
&= \inf\left\{\alpha \in (0,1), F_T\left(|z_{obs}|\right) \geq 1-\alpha/2\right\} \\
&= 2 - 2F_T\left(|z_{obs}|\right) \\
&= \mathbb{P}\left[|T| \geq |z_{obs}|\right).
\end{aligned}$$

Exemple : Cet exemple est tiré du cours de B. Portier à l'INSA de Rouen. On s'intéresse à un échantillon de 30 matières fécales. On soumet cet échantillon à deux méthodes différentes de spectrométrie pour étudier leur teneur en lutécium radioactif. Les mesures relatives aux deux méthodes ne

peuvent donc bien évidemment pas être indépendantes. On souhaite savoir si les résultats obtenus avec ces deux méthodes sont équivalents, i.e si les moyennes sont les mêmes. On obtient

$$\overline{x}_{30} = 120.83 \qquad \text{et} \qquad \overline{y}_{30} = 119.33.$$

En effectuant le test pour les données appariées, on obtient une p-value égale à 0.66 et au risque de 5%, on ne rejette donc pas H0. A noter que si on avait effectué le test de Student pour des variables indépendantes, on aurait obtenue un p-value égale à 0.0031 et on aurait alors rejeter H0. Cette différence est due à la forte corrélation entre les données. En effet le coefficient de corrélation entre les données est de 0.982.

5.5.3 Test d'inégalité $\mu_1 \leq \mu_2$

Le test : On teste au risque $\alpha \in (0,1)$ l'hypothèse nulle H0 : "$\mu_1 \leq \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 > \mu_2$". Sous H0, il existe $\mu' \leq 0$ tel que

$$Z(\mu') = \sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - \mu'}{S} \sim T_{n-1}.$$

On définit alors la zone de rejet par

$$ZR = \left\{ Z(0) > t_{n-1,1-\alpha} \right\},$$

où $t_{n-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté. On calcule $z_{obs} = \sqrt{n}\frac{\overline{x}_n - \overline{y}_n}{s}$. Si $z_{obs} > t_{n-1,1-\alpha}$, on rejette H0 et inversement.

Remarque 1 : En réalité, pour trouver la zone de rejet, comme $\mu = \mu_1 - \mu_2$ peut être aussi petit que l'on veut sous H0, on doit rejeter H0 si $\overline{X}_n - \overline{Y}_n$ est trop grand, i.e on cherche c_α tel que

$$\sup_{\mu \leq 0} \mathbb{P}_\mu \left["\text{On rejette H0"}\right] = \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\overline{X}_n - \overline{Y}_n \geq c_\alpha\right] = \alpha.$$

Or, si on note F_T la fonction de répartition de la loi de Student à $n-1$ degrés de liberté, on obtient

$$\begin{aligned}
\alpha &= \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\overline{X}_n - \overline{Y}_n \geq c_\alpha\right] \\
&= \sup_{\mu \leq 0} \mathbb{P}_\mu \left[\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - \mu}{S} \geq \sqrt{n}\frac{c_\alpha - \mu}{S}\right] \\
&= \sup_{\mu \leq 0} \left(1 - F_T\left(\sqrt{n}\frac{c_\alpha - \mu}{S}\right)\right).
\end{aligned}$$

Comme la fonction de répartition F_T est strictement croissante, on obtient

$$\alpha = 1 - F_T\left(\frac{c_\alpha}{\sqrt{n}S}\right),$$

et donc en appliquant F_T^{-1},

$$t_{n-1,1-\alpha} = \frac{c_\alpha}{\sqrt{n}S},$$

et donc

$$c_\alpha = t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}.$$

On retrouve ainsi la zone de rejet, i.e

$$ZR = \left\{\overline{X}_n - \overline{Y}_n \geq t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}\right\} = \left\{\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n}{S} \geq t_{n-1,1-\alpha}\right\}.$$

Remarque 2 : On a

$$p - \text{value} = \mathbb{P}\left[T \geq z_{obs}\right],$$

où $T \sim T_{n-1}$. En effet ;

$$\begin{aligned} p - \text{value} &= \inf\left\{\alpha \in (0,1), \text{"On rejette H0"}\right\} \\ &= \inf\left\{\alpha \in (0,1), z_{obs} \geq t_{n-1,1-\alpha}\right\} \\ &= \inf\left\{\alpha \in (0,1), F_T\left(z_{obs}\right) \geq 1-\alpha\right\} \\ &= 1 - F_T\left(z_{obs}\right). \end{aligned}$$

5.5.4 Test d'inégalité $\mu_1 \geq \mu_2$

On veut tester au risque $\alpha \in (0,1)$ l'hypothèse nulle H0 : "$\mu_1 \geq \mu_2$" contre l'hypothèse alternative H1 : "$\mu_1 < \mu_2$". Sous H0, il existe $\mu' \geq 0$ tel que

$$Z\left(\mu'\right) = \sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - \mu'}{S} \sim T_{n-1}.$$

On définit alors la zone de rejet par

$$ZR = \left\{Z(0) < -t_{n-1,1-\alpha}\right\},$$

où $t_{n-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Student à $n-1$ degrés de liberté. On calcul $z_{obs} = \sqrt{n}\frac{\overline{x}_n - \overline{y}_n}{s}$. Si $z_{obs} < -t_{n-1,1-\alpha}$, on rejette H0 et inversement.

Remarque 1 : En réalité, pour trouver la zone de rejet, comme $\mu = \mu_1 - \mu_2$ peut être aussi grand que l'on veut sous H0, on doit rejeter H0 si $\overline{X}_n - \overline{Y}_n$ est trop petit, i.e on cherche c_α tel que

$$\sup_{\mu \geq 0} \mathbb{P}_\mu\left[\text{"On rejette H0"}\right] = \sup_{\mu \geq 0} \mathbb{P}_\mu\left[\overline{X}_n - \overline{Y}_n \leq c_\alpha\right] = \alpha.$$

Or si on note F_T la fonction de répartition de la loi de Student à $n-1$ degrés de liberté, on obtient

$$\begin{aligned}\alpha &= \sup_{\mu\geq 0} \mathbb{P}_\mu\left[\overline{X}_n - \overline{Y}_n \leq c_\alpha\right] \\ &= \sup_{\mu\geq 0} \mathbb{P}_\mu\left[\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n - \mu}{S} \leq \sqrt{n}\frac{c_\alpha - \mu}{S}\right] \\ &= \sup_{\mu\geq 0} F_T\left(\sqrt{n}\frac{c_\alpha - \mu}{S}\right)\end{aligned}$$

Comme la fonction F_T est strictement croissante, on obtient

$$\alpha = F_T\left(\sqrt{n}\frac{c_\alpha}{S}\right),$$

et donc, en appliquant F_T^{-1},

$$t_{n-1,\alpha} = \sqrt{n}\frac{c_\alpha}{S}.$$

Ainsi, comme $t_{n-1,\alpha} = -t_{n-1,1-\alpha}$,

$$c_\alpha = -t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}.$$

On retrouve ainsi la zone de rejet, i.e

$$ZR = \left\{\overline{X}_n - \overline{Y}_n \leq -t_{n-1,1-\alpha}\frac{S}{\sqrt{n}}\right\} = \left\{\sqrt{n}\frac{\overline{X}_n - \overline{Y}_n}{S} \leq -t_{n-1,1-\alpha}\right\}.$$

Remarque 2 : On a

$$p - \text{value} = \mathbb{P}\left[T \leq z_{obs}\right],$$

où $T \sim T_{n-1}$. En effet,

$$\begin{aligned}p - \text{value} &= \inf\left\{\alpha \in (0,1), \text{"On rejette H0"}\right\} \\ &= \inf\left\{\alpha \in (0,1), z_{obs} \leq -t_{n-1,1-\alpha}\right\} \\ &= \inf\left\{\alpha \in (0,1), F_T\left(z_{obs}\right) \leq \alpha\right\} \\ &= F_T\left(z_{obs}\right).\end{aligned}$$

5.6 Tests du Khi-deux

5.6.1 Test d'indépendance du Khi-deux

Le test d'indépendance du Khi-deux permet de vérifier l'indépendance entre deux variables aléatoires X, Y, i.e on veut tester au risque α l'hypothèse nulle H0 : "X et Y sont indépendantes" contre l'hypothèse alternative H1 : "X et Y ne sont pas indépendantes".

Le cadre : On dispose de n données $(x_1, y_1), \ldots, (x_n, y_n)$ qui sont des réalisations d'un couple de

variables aléatoires (X, Y) à valeurs dans $\{a_1, \ldots, a_p\} \times \{b_1, \ldots, b_q\}$.

Construction du test : Le test du Khi-deux repose sur le fait que les variables aléatoires X et Y sont indépendantes si et seulement si pour tout $(i, j) \in \{1, \ldots, p\} \times \{1, \ldots, q\}$,

$$\mathbb{P}\left[X = a_i \text{ et } Y = b_j\right] = \mathbb{P}\left[X = a_i\right] \mathbb{P}\left[Y = b_j\right].$$

Pour tester si les variables aléatoires X et Y sont indépendantes, il suffit donc de comparer les estimations des probabilités jointes au produit des estimations des probabilités marginales.

Notations : Les réalisations $(x_1, y_1), \ldots, (x_n, y_n)$ sont des réalisations des couples de variables aléatoires (X_k, Y_k), où les (X_k, Y_k) sont indépendants et de même loi que (X, Y). Pour tout $(i, j) \in \{1, \ldots, p\} \times \{1, \ldots, q\}$, on note :

- $O_{i,j} = \sum_{k=1}^{n} \mathbf{1}_{\{X_k = a_i, Y_k = b_j\}}$ qui désigne le nombre de données pour lesquelles $(X, Y) = (a_i, b_j)$.
- $O_{i,\cdot} = \sum_{j=1}^{q} O_{i,j}$ qui désigne le nombre de données pour lesquelles $X = a_i$.
- $O_{\cdot,j} = \sum_{i=1}^{p} O_{i,j}$ qui désigne le nombre de données pour lesquelles $Y = b_j$.
- $E_{i,j} = \frac{O_{i,\cdot} \times O_{\cdot,j}}{n}$.

Notons que

- $O_{i,j}/n$ est un estimateur de $\mathbb{P}\left[X = a_i, Y = b_j\right]$.
- $E_{i,j}/n$ est un estimateur de $\mathbb{P}\left[X = a_i\right] \mathbb{P}\left[Y = b_j\right]$.

Sous l'hypothèse H0, ces deux estimateurs sont censés être proches. On introduit la variable

$$Z = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}.$$

Proposition 5.6.1 (Admise)**.** *Si X et Y sont indépendantes, alors,*

$$Z \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2_{(p-1)(q-1)}.$$

La proposition précédente est fausse sous $H1$.

Le test :

- On teste au risque α l'hypothèse nulle H0 : "X et Y sont indépendantes" contre l'hypothèse alternative H1 : "X et Y ne sont pas indépendantes".
- On a la statistique de test

$$Z = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2_{(p-1)(q-1)} \quad \text{sous H0.}$$

- On a la zone de rejet

$$ZR = \left\{Z > k_{(p-1)(q-1),1-\alpha}\right\}$$

où $k_{(p-1)(q-1),1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi du Khi-deux à $(p-1)(q-1)$ degrés de liberté.

— Si z_{obs} appartient à la zone de rejet, alors on rejette H0 et inversement.

Exemple : Cet exemple est issu du cours de B. Portier à l'INSA de Rouen. On étudie l'influence du sexe sur la couleur des cheveux d'un groupe d'élèves. On veut savoir si la couleur des cheveux est indépendante du sexe. Pour cela, on dispose des données suivantes :

Sexe	Blond	Roux	Châtain	Brun	Noir de Jais	Total
Garçon	592	119	849	504	36	2100
Fille	544	97	677	451	14	1783
Total	1136	2126	1526	955	50	3883

On désigne par X la variable couleur des cheveux et par Y la variable sexe.

On teste au risque de 5% l'hypothèse H0 : "X et Y sont indépendantes" contre H1 : "X et Y ne sont pas indépendantes". On a la statistique de test

$$Z = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_4 \quad \text{sous H0.}$$

On a la zone de rejet

$$ZR = \{Z > 9.50\}.$$

Ici, $z_{obs} = 10.47$ et on rejette donc l'hypothèse H0. A noter que l'on a une p-value égale à 0.033 et on n'aurait donc pas rejeté H0 au risque de 1%.

5.6.2 Test d'adéquation

On dispose de n données $x_1, \ldots, x_n$ qui sont les réalisations de variables aléatoires indépendantes $X_1, ..., X_n$ et de même loi, à valeurs dans $\{a_1, \ldots, a_K\}$. On note P la loi inconnue de la variable aléatoire discrète X et on souhaite savoir si $P = P0$ où $P0$ est une loi connue. On veut donc tester au risque α l'hypothèse nulle H0 : "la loi de X est $P0$" i.e "$P = P0$" contre l'hypothèse alternative H1 : "la loi de X n'est pas P0", i.e "$P \neq P0$".

Notations : On note $p_{0,1}, ..., p_{0,K} > 0$ les probabilités définissant la loi P0. On note également pour tout $k = 1, \ldots, K$:

— E_k l'effectif observé pour la modalité a_k, i.e

$$E_k = \sum_{i=1}^{n} \mathbf{1}_{\{X_j = a_k\}}.$$

— N_k l'effectif théorique pour la modalité a_k sous la loi P0, i.e

$$N_k = np_{0,k}.$$

Construction de la statistique de test : La variable E_k/n est donc un estimateur naturel de $p_k = \mathbb{P}[X = a_k]$. Sous l'hypothèse nulle, cet estimateur doit donc converger vers $N_k/n = p_{0,k}$. On cherche alors une statistique de test et on s'intéresse pour cela à

$$Q^2 = \sum_{k=1}^{K} \frac{(E_k - N_k)^2}{N_k}.$$

Proposition 5.6.2. *Si la loi de X est P0, alors*

$$Q^2 \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_{K-1}$$

Démonstration. A noter que le TLC pour chaque E_k nous donne

$$\sqrt{n}\left(\frac{E_k}{n} - p_k\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, p_k(1-p_k))$$

mais il est inexploitable pour obtenir le résultat. On va plutôt utiliser le théorème de Cochran. Pour cela, on remarque que l'on peut écrire

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Y_{1,i} \\ \vdots \\ Y_{K,i} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} E_1 \\ \vdots \\ E_K \end{pmatrix}$$

avec $Y_{k,i} = \mathbf{1}_{X_i=k}$. Les vecteurs Y_i sont i.i.d avec

$$\mathbb{E}[Y_1] = \begin{pmatrix} p_1 \\ \vdots \\ p_K \end{pmatrix}$$

et comme si $k \neq k'$, on a $Y_{i,k}Y_{i,k'} = 0$, on obtient

$$\mathrm{Var}[Y_1] = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots \\ -p_1p_2 & p_2(1-p_2) & \dots \\ \vdots & \vdots & \vdots \end{pmatrix}$$

Le TLC multidimensionnel nous donne donc

$$\sqrt{n}\left(\overline{Y}_n - \mathbb{E}[Y_1]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}[Y_1])$$

et comme $\mathrm{Var}[Y_1] = \mathrm{diag}\left(\sqrt{p_1}, \dots, \sqrt{p_K}\right)\left(I - \sqrt{p}\sqrt{p}^T\right)\mathrm{diag}\left(\sqrt{p_1}, \dots, \sqrt{p_K}\right)$, on peut donc ré-

écrire le TLC comme

$$\sqrt{n}\begin{pmatrix}\frac{\overline{Y}_{1,n}-p_1}{\sqrt{p_1}}\\ \vdots \\ \frac{\overline{Y}_{K,n}-p_1}{\sqrt{p_K}}\end{pmatrix} = \sqrt{n}\begin{pmatrix}\frac{E_1/n-p_1}{\sqrt{p_1}}\\ \vdots \\ \frac{E_K/n-p_1}{\sqrt{p_K}}\end{pmatrix} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(0, I-\sqrt{p}\sqrt{p}^T\right)$$

Or $\sqrt{p}\sqrt{p}^T$ est la matrice de projection orthogonale sur $V = \text{Vect}\{p\}$, et $I-\sqrt{p}\sqrt{p}^T$ est la projection orthogonale sur $V^\perp$. Et on a donc, en appliquant le théorème de continuité et le théorème de Cochran

$$\sum_{k=1}^{K}\frac{(E_k-N_k)^2}{N_k} = n\left\|\begin{pmatrix}\frac{E_1/n-p_1}{\sqrt{p_1}}\\ \vdots \\ \frac{E_K/n-p_1}{\sqrt{p_K}}\end{pmatrix}\right\|^2 \xrightarrow[n\to\infty]{\mathcal{L}} \left\|\mathcal{N}\left(0, I-\sqrt{p}\sqrt{p}^T\right)\right\|^2 \sim \chi^2_{K-1}.$$

□

La proposition précédente est fausse si l'hypothèse H0 n'est pas vérifiée.

Le test

— On test au risque α l'hypothèse nulle H0 : "X suit la loi P0" contre H1 : "X ne suit pas la loi P0".

— On a la statistique de test

$$Q^2 = \sum_{k=1}^{K}\frac{(E_k-N_k)^2}{N_k} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_{K-1} \quad \text{sous H0}$$

— On a la zone de rejet

$$ZR = \left\{Q^2 > k_{1-\alpha}\right\}$$

où $k_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi du Khi-deux à $K-1$ degrés de liberté.

— Si q^2_{obs} appartient à la zone de rejet, on rejette H0 et inversement.

Exemple : Les données suivantes sont issues du cours de B. Portier à l'INSA de Rouen. Un biologiste pense qu'à un âge donné :

— 50% des bébés marchent

— 12% ont une ébauche de marche

— 38% ne marchent pas

On fait une étude sur 80 bébés et on obtient :

— 35 de ces bébés ne marchent pas

— 4 ont une ébauche de marche

— 41 ne marchent pas

On souhaite donc vérifier que le biologiste a tort. On résume les données dans le tableau suivant :

	Marche	Ebauche	Ne marche pas	Total
Effectif observé	35	4	41	80
Effectif Théorique	$80\times$ 0.5	$80\times$ 0.12	$80\times$ 0.38	80
"Distance"	0.625	3.267	3.696	7.588

On teste donc au risque 5% l'hypothèse nulle H0 : "X suit la loi P0" contre H1 : "X ne suit pas la loi P0". On a la statistique de test

$$Q^2 = \sum_{k=1}^{3} \frac{(E_k - N_k)^2}{N_k} \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_2 \quad \text{sous H0}$$

On a la zone de rejet

$$ZR = \left\{Q^2 > 5.99\right\}.$$

Ici, $q^2_{obs} = 7.5878$ et on rejette donc H0.

5.7 Tests asymptotiques

5.7.1 Introduction

On considère ici des données $x_1, \ldots, x_n$ qui sont des réalisations de variables aléatoires $X_1, \ldots, X_n$ qui sont indépendantes et de même loi que X. On s'intéresse à l'estimation d'un paramètre θ de la loi de X, et on suppose que l'on a un estimateur $\hat{\theta}_n$ asymptotiquement normal, i.e qu'il existe $\sigma^2 > 0$ tel que

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

On suppose également que l'on dispose d'un estimateur consistant $\hat{\sigma}^2_n$ de la variance asymptotique σ^2. A l'aide du théorème de Slutsky, on a alors

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

5.7.2 Test d'égalité

On veut tester au risque α l'hypothèse nulle H0 : "$\theta = \theta_0$" contre l'hypothèse alternative H1 : "$\theta \neq \theta_0$". On a la statistique de test

$$Z = \sqrt{n}\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \quad \text{sous H0}$$

On obtient alors la zone de rejet

$$ZR = \left\{|z_{obs}| > q_{1-\alpha/2}\right\},$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite. Si z_{obs} appartient à la zone de rejet, on rejette H0 et inversement.

Remarque 1 : Cela revient à vérifier que θ_0 appartient à l'intervalle de confiance asymptotique de niveau $1-\alpha$ de θ. En effet,

$$\begin{aligned}|z_{obs}| \leq q_{1-\alpha/2} &\Leftrightarrow -q_{1-\alpha/2} \leq \sqrt{n}\frac{\theta_n - \theta_0}{\sigma_n} \leq q_{1-\alpha/2} \\ &\Leftrightarrow \theta_n - q_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}} \leq \theta_0 \leq \theta_n + q_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}} \\ &\Leftrightarrow \theta_0 \in IC_{1-\alpha}(\theta)\end{aligned}$$

Remarque 2 : En réalité, pour trouver la zone de rejet, comme $\hat{\theta}_n$ doit être proche de θ, on cherche c_α tel que

$$\lim_{n\to\infty} \mathbb{P}_{\theta=\theta_0}\left["\text{On rejette H0"}\right] = \lim_{n\to\infty} \mathbb{P}\left[\left|\hat{\theta}_n - \theta_0\right| \geq c_\alpha\right] = \alpha.$$

On obtient donc

$$\alpha = \lim_{n\to\infty} \mathbb{P}_{\theta=\theta_0}\left[\left|\hat{\theta}_n - \theta_0\right| \geq c_\alpha\right] = \lim_{n\to\infty} 2\mathbb{P}_{\theta=\theta_0}\left[\sqrt{n}\frac{\hat{\theta}_n - \theta_0}{\hat{\sigma}_n} \geq \frac{\sqrt{n}}{\hat{\sigma}_n}c_\alpha\right].$$

Comme

$$\sqrt{n}\frac{\hat{\theta}_n - \theta}{\hat{\sigma}_n} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

on choisit donc

$$\frac{\sqrt{n}}{\hat{\sigma}_n}c_\alpha = q_{1-\alpha/2} \Leftrightarrow c_\alpha = q_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}.$$

On retrouve alors la zone de rejet

$$ZR = \left\{\left|\hat{\theta}_n - \theta_0\right| \geq q_{1-\alpha/2}\frac{\hat{\sigma}_n}{\sqrt{n}}\right\} = \left\{\sqrt{n}\frac{\left|\hat{\theta}_n - \theta_0\right|}{\hat{\sigma}_n} \geq q_{1-\alpha/2}\right\}.$$

Remarque 3 : Soit $Z \sim \mathcal{N}(0,1)$, on a alors

$$p - \text{value} = \mathbb{P}\left[|Z| \geq |z_{obs}|\right].$$

En effet,

$$\begin{aligned}p - \text{value} &= \inf\left\{\alpha \in (0,1), "\text{On rejette H0"}\right\} \\ &= \inf\left\{\alpha \in (0,1), |z_{obs}| \geq q_{1-\alpha/2}\right\} \\ &= \inf\left\{\alpha \in (0,1), F\left(|z_{obs}|\right) \geq 1-\alpha/2\right\} \\ &= 2 - 2F\left(|z_{obs}|\right) \\ &= 2\mathbb{P}\left[Z \geq |z_{obs}|\right].\end{aligned}$$

Par symétrie de la loi normale,

$$p - \text{value} = 2\mathbb{P}\left[Z \geq |z_{obs}|\right] = \mathbb{P}\left[Z \geq |z_{obs}|\right] + \mathbb{P}\left[-Z \geq |z_{obs}|\right] = \mathbb{P}\left[|Z| \geq |z_{obs}|\right].$$

5.7.3 Test d'inégalité $\theta_0 \geq \theta$

On veut tester au risque α l'hypothèse nulle H0 : "$\theta_0 \geq \theta$" contre l'hypothèse alternative H1 : "$\theta_0 < \theta$". Sous H0, il existe $\theta' \leq \theta_0$ tel que

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

On définit alors la zone de rejet

$$ZR = \left\{Z\left(\theta_0\right) > q_{1-\alpha}\right\}$$

où $q_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi normale centrée réduite. On calcule $z_{obs} = \sqrt{n}\frac{\theta_n - \theta_0}{\sigma_n}$, et si $z_{obs} > q_{1-\alpha}$ on rejette H0 et inversement.

Remarque : On a

$$p - \text{value} = \mathbb{P}\left[Z \geq z_{obs}\right],$$

où $Z \sim \mathcal{N}(0,1)$. En effet,

$$\begin{aligned} p - \text{value} &= \inf\left\{\alpha \in (0,1), \text{"On rejette H0"}\right\} \\ &= \inf\left\{\alpha \in (0,1), z_{obs} \geq q_{1-\alpha}\right\} \\ &= \inf\left\{\alpha \in (0,1), F\left(z_{obs}\right) \geq 1-\alpha\right\} \\ &= 1 - F\left(z_{obs}\right). \end{aligned}$$

5.7.4 Test d'inégalité $\theta \geq \theta_0$

On veut tester au risque $\alpha \in (0,1)$, l'hypothèse nulle H0 : "$\theta \geq \theta_0$" contre l'hypothèse alternative H1 : "$\theta < \theta_0$". Sous H0, il existe $\theta' \geq \theta_0$ tel que

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

On définit alors la zone de rejet par

$$ZR = \left\{Z\left(\theta_0\right) < -q_{1-\alpha}\right\},$$

où q_α est le quantile d'ordre α de la loi normale centrée réduite. Si $z_{obs} < q_\alpha$, alors on rejette H0 et inversement.

Remarque : On a

$$p - \text{value} = \mathbb{P}\left[Z \leq z_{obs}\right],$$

où Z suit une loi normale centrée réduite. En effet,

$$\begin{aligned} p-\text{value} &= \inf\{\alpha \in (0,1), \text{"On rejette H0"}\} \\ &= \inf\{\alpha \in (0,1), z_{obs} \leq q_\alpha\} \\ &= \inf\{\alpha \in (0,1), F(z_{obs}) \leq \alpha\} \\ &= F(z_{obs}). \end{aligned}$$