

MAP 433 : Introduction aux méthodes statistiques

27 août 2015

Organisation : équipe enseignante

Cours

Eric MOULINES, Ecole Polytechnique
eric.moulines@polytechnique.edu

PC

- ▶ Gersende Fort, DR CNRS, Télécom ParisTech,
- ▶ Lucas Gérin, École Polytechnique,
- ▶ Christophe Giraud, Université Paris-Sud et École Polytechnique,
- ▶ Marc Lavielle, DR INRIA, INRIA Saclay,
- ▶ Matthieu Lerasle, CR CNRS, Université de Nice,
- ▶ Mathieu Rosenbaum, Université Pierre-et-Marie Curie,
- ▶ Francois Roueff, Professeur, Télécom ParisTech.

Organisation : materiel

- ▶ **Transparents** du cours téléchargeables à l'adresse
[https ://moodle.polytechnique.fr/course/view.php?id=1717](https://moodle.polytechnique.fr/course/view.php?id=1717)
- ▶ **Polycopié document autonome contenant l'intégralité du cours et plus, téléchargeable à la même adresse.**
- ▶ Les documents et **exercices** de PC. [les exercices obligatoires et pour aller plus loin]
- ▶ Des liens pour des expériences numériques.

Présentation (succincte) du cours

- ▶ Introduction aux statistiques et rappels de probabilités (1 cours)
- ▶ Expérience statistique (1 cours).
- ▶ Méthodes d'estimation classique (2 cours).
- ▶ Information statistique, théorie asymptotique pour l'estimation (2 cours).
- ▶ Décision statistique et tests (2 cours).

Plan

- ▶ **Problématique statistique** : de quoi s'agit-il ?
- ▶ **Echantillonnage**.
- ▶ Estimation d'une distribution inconnue à partir d'un n -échantillon, **méthodes empiriques**.

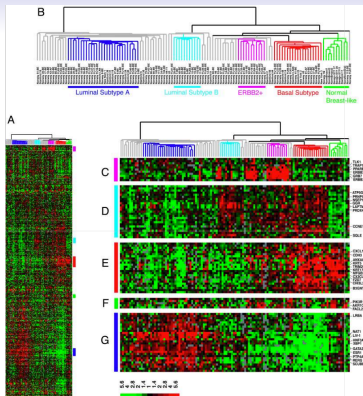
E-medicine



6 / 29

FIGURE – Identifier les facteurs de risque pour le développement d'un cancer

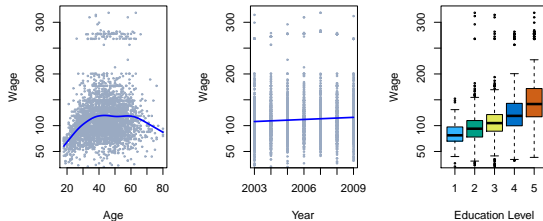
Génomique-Protéomique



17 / 29

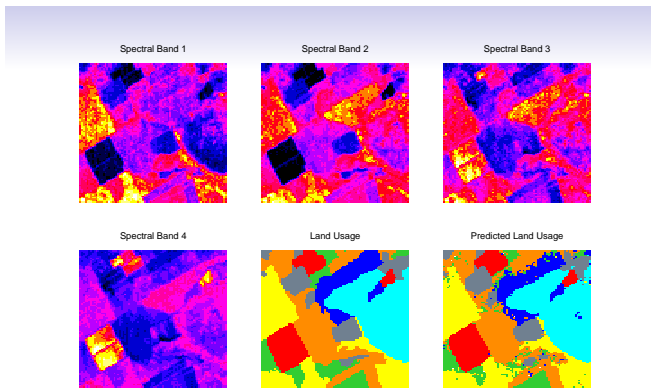
FIGURE – classifier des tissus en fonctions de données d'expression de gènes

Économie



Income survey data for males from the central Atlantic region of the USA in 2009.

Télé-détection



Usage $\in \{\text{red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil}\}$

Et plein d'autres choses

- ▶ Finance
- ▶ Opinion
- ▶ Marketing
- ▶ Sport

Problématique statistique

- **Point de départ** : des observations (des nombres réels)

$$x_1, \dots, x_n.$$

- **Modélisation statistique** :

- les observations sont des réalisations

$$X_1(\omega), \dots, X_n(\omega) \text{ de v.a.r. } X_1, \dots, X_n.$$

- La loi $\mathbb{P}^{(X_1, \dots, X_n)}$ de (X_1, \dots, X_n) **est inconnue**, mais appartient à une famille donnée

$$\{ \mathbb{P}_{\vartheta}^n, \vartheta \in \Theta \}.$$

- **Problématique** : à partir de « l'observation » x_1, \dots, x_n , peut-on **retrouver** \mathbb{P}_{ϑ}^n ? et donc ϑ ?

Problématique statistique (suite)

- ▶ ϑ est le **paramètre** et Θ l'**ensemble** des paramètres.
- ▶ **Estimation** : à partir de X_1, \dots, X_n , construire $\varphi_n(X_1, \dots, X_n)$ qui « approche au mieux » ϑ .
- ▶ **Test** : à partir de X_1, \dots, X_n , établir une **décision** $\varphi_n(X_1, \dots, X_n) \in \{\text{ensemble de décisions}\}$ concernant ϑ pouvant être vraie ou fausse.

Exemple le plus simple

- ▶ On lance une pièce de monnaie 18 fois et on observe ($P = 0$, $F = 1$)

0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0.

- ▶ Modèle statistique : on observe $n = 18$ variables aléatoires X_i indépendantes, de Bernoulli de paramètre **inconnu** $\vartheta \in \Theta = [0, 1]$.
 - ▶ **Estimation**. Estimateur $\bar{X}_{18} = \frac{1}{18} \sum_{i=1}^{18} X_i \stackrel{\text{ici}}{=} 8/18 = 0.44$.
Quelle précision ?
 - ▶ **Test**. Décision à prendre : « la pièce est-elle équilibrée » ?. Par exemple : on compare \bar{X}_{18} à 0.5. Si $|\bar{X}_{18} - 0.5|$ « petit », on accepte l'hypothèse « la pièce est équilibrée ». Sinon, on rejette. Quel seuil choisir, et avec quelles conséquences (ex. probabilité de se tromper).

Echantillonnage

- ▶ L'expérience statistique la plus centrale : on observe la réalisation de X_1, \dots, X_n , v.a.r. où les X_i sont **indépendantes, identiquement distribuées**, de même loi commune \mathbb{P}^X .
- ▶ Que dire de la loi \mathbb{P}^X commune des X_i ?
- ▶ Structure stochastique **très simple** (variable aléatoires indépendantes, de même loi). Mais : espace des paramètres **immense** (toutes les lois de probabilités).

Rappel : loi d'une variable aléatoire réelle

Definition

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \longrightarrow (\mathbb{R}, \mathcal{B})$$

Loi de X : mesure de probabilité sur $(\mathbb{R}, \mathcal{B})$, notée \mathbb{P}^X , définie par

$$\mathbb{P}^X [A] = \mathbb{P} [X^{-1}(A)], \quad A \in \mathcal{B}.$$

Formule d'intégration

$$\mathbb{E} [\varphi(X)] = \int_{\Omega} \varphi(X(\omega)) \mathbb{P}(d\omega) = \int_{\mathbb{R}} \varphi(x) \mathbb{P}^X(dx)$$

φ fonction test.

Loi d'une variable aléatoire (suite)

Exemple 1 : X suit la loi de Bernoulli de paramètre $1/3$.

- ▶ La loi de X est décrite par

$$\mathbb{P}[X = 1] = \frac{1}{3} = 1 - \mathbb{P}[X = 0].$$

- ▶ Ecriture de $\mathbb{P}^X(dx)$:

$$\mathbb{P}^X(dx) = \frac{1}{3}\delta_1(dx) + \frac{2}{3}\delta_0(dx).$$

- ▶ Formule de calcul

$$\begin{aligned}\mathbb{E}[\varphi(X)] &= \int_{\mathbb{R}} \varphi(x) \mathbb{P}^X(dx) \\ &= \frac{1}{3} \int_{\mathbb{R}} \varphi(x) \delta_1(dx) + \frac{2}{3} \int_{\mathbb{R}} \varphi(x) \delta_0(dx) \\ &= \frac{1}{3} \varphi(1) + \frac{2}{3} \varphi(0).\end{aligned}$$

Loi d'une variable aléatoire (suite)

Exemple 2 : $X \sim$ loi de Poisson de paramètre 2.

- ▶ La loi de X est décrite par

$$\mathbb{P}[X = k] = e^{-2} \frac{2^k}{k!}, \quad k = 0, 1, \dots$$

- ▶ Ecriture de $\mathbb{P}^X(dx)$:

$$\mathbb{P}^X(dx) = e^{-2} \sum_{k \in \mathbb{N}} \frac{2^k}{k!} \delta_k(dx).$$

- ▶ Formule de calcul

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x) \mathbb{P}^X(dx) = e^{-2} \sum_{k \in \mathbb{N}} \varphi(k) \frac{2^k}{k!}.$$

Loi d'une variable aléatoire (suite)

Exemple 3 : $X \sim \mathcal{N}(0, 1)$ (loi normale standard).

- ▶ La loi de X est décrite par

$$\mathbb{P}[X \in [a, b]] = \int_{[a, b]} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}$$

- ▶ Ecriture de $\mathbb{P}^X(dx)$:

$$\mathbb{P}^X(dx) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

dx : mesure de Lebesgue.

- ▶ Formule de calcul

$$\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x) \mathbb{P}^X(dx) = \int_{\mathbb{R}} \varphi(x) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}}.$$

Loi d'une variable aléatoire (suite)

Exemple 4 : $X = Z \wedge 1$, où la loi de Z a une densité f par rapport à la mesure de Lebesgue sur \mathbb{R} .

Loi de X

- ▶ Sur l'événement $\{Z < 1\}$, on observe $X = Z$.
- ▶ Sur l'événement $\{Z \geq 1\}$, on observe $X = 1$.

Ecriture de $\mathbb{P}^X(dx)$:

$$\mathbb{P}^X(dx) = f(x)1_{\{x < 1\}} dx + \mathbb{P}[Z \geq 1]\delta_1(dx),$$

c'est-à-dire

$$\mathbb{P}^X(dx) = f(x)1_{\{x < 1\}} dx + \left(\int_{[1, +\infty)} f(u) du \right) \delta_1(dx)$$

Formule de calcul

$$\begin{aligned} \mathbb{E} [\varphi(X)] &= \int_{\mathbb{R}} \varphi(x) \mathbb{P}^X(dx) \\ &= \int_{(-\infty, 1)} \varphi(x) f(x) dx + \left(\int_{[1, +\infty)} f(u) du \right) \varphi(1). \end{aligned}$$

Identification de la loi : fonction de répartition

- ▶ La loi d'une variable aléatoire X est un « objet compliqué » :
 - ▶ elle peut être discrète (somme de masses de Dirac)
 - ▶ elle peut être (absolument) continue (densité par rapport à la mesure de Lebesgue)
 - ▶ elle peut-être une combinaison des deux, ou encore autre chose....
- ▶ On peut **caractériser la loi** de X par un objet plus simple à manipuler : une fonction croissante bornée : la **fonction de répartition**.
- ▶ Plus facile à étudier dans un **contexte de statistique**.
- ▶ (Il y aura bien sûr des limites à cette approche...)

Fonction de répartition

Definition

X variable aléatoire réelle. Fonction de répartition de X :

$$F(x) := \mathbb{P} [X \leq x], \quad x \in \mathbb{R}.$$

- ▶ F est croissante, cont. à droite, $F(-\infty) = 0$, $F(+\infty) = 1$
- ▶ F caractérise la loi \mathbb{P}^X :

$$\mathbb{P}^X [(a, b)] = \mathbb{P} [a < X \leq b] = F(b) - F(a)$$

- ▶ Désormais, la loi (distribution) de X désignera indifféremment F ou \mathbb{P}^X .

Problématique statistique

- ▶ On « observe »

$$X_1, \dots, X_n \sim_{i.i.d.} F,$$

F fonction de répartition **quelconque, inconnue**.

- ▶ Terminologie : (X_1, \dots, X_n) est un **n -échantillon** de la loi F .
- ▶ Comment **retrouver** F à partir des observations X_1, \dots, X_n ?
- ▶ **Démarche** : on construit une fonction (aléatoire)

$x \rightsquigarrow \hat{F}_n(x) = F_n(x; X_1, \dots, X_n)$ ne dépendant pas de F (inconnu) telle que

$$\hat{F}_n(x) - F(x)$$

petit lorsque n grand... Comment ? Petit dans quel sens ?

Fonction de répartition empirique

Definition

Fonction de répartition empirique associée au n -échantillon (X_1, \dots, X_n) :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{X_i \leq x\}}, \quad x \in \mathbb{R}.$$

- ▶ C'est une fréquence empirique
- ▶ Terminologie : \hat{F}_n est un **estimateur** : fonction des observations qui ne dépend **pas** de la quantité inconnue.
- ▶ Pour tout $x_0 \in \mathbb{R}$

$$\hat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0), \quad n \rightarrow \infty$$

(loi faible des grands nombres appliquée aux $1_{\{X_i \leq x_0\}}$).

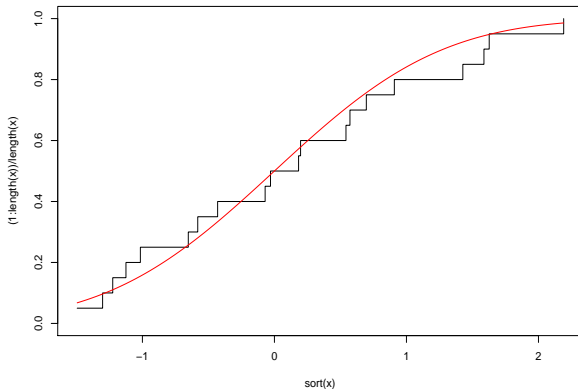


FIGURE – \hat{F}_n (noir), F (rouge), $n = 20$. $F \sim \mathcal{N}(0, 1)$.

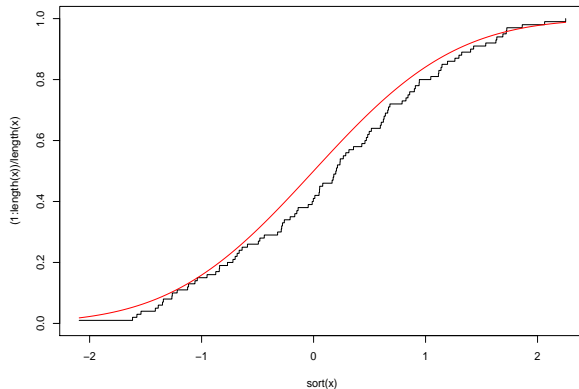


FIGURE – \hat{F}_n (noir), F (rouge), $n = 100$. $F \sim \mathcal{N}(0, 1)$.

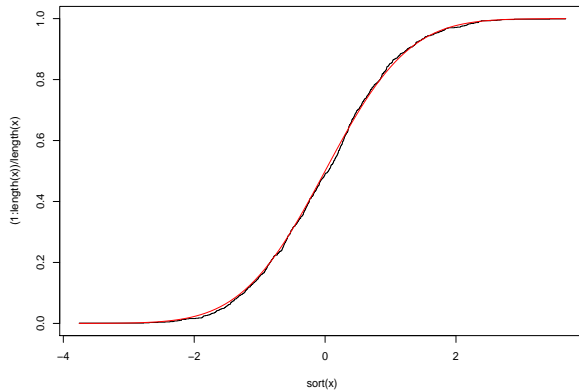


FIGURE – \hat{F}_n (noir), F (rouge), $n = 1000$. $F \sim \mathcal{N}(0,1)$.

Convergence en probabilité

- ▶ Mode de convergence « naturel » en statistique

- ▶ **Rappel** : $X_n \xrightarrow{\mathbb{P}} X$ si

$$\forall \varepsilon > 0, \mathbb{P} [|X_n - X| \geq \varepsilon] \rightarrow 0, \quad n \rightarrow \infty.$$

- ▶ **Interprétation** : pour tout niveau de risque $\alpha > 0$ (petit) et tout niveau de précision $\varepsilon > 0$, il existe un rang $N = N(\alpha, \varepsilon)$ tel que

$$n > N \text{ implique } |X_n - X| \leq \varepsilon \text{ avec proba. } \geq 1 - \alpha.$$

- ▶ En pratique, on souhaite simultanément N , α et ε petits. Quantités **antagonistes** (à suivre...).

Vers la précision d'estimation

- ▶ On a $\forall x_0 \in \mathbb{R}, \hat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$. Avec **quelle précision** ?
Problèmes de même types :
 - ▶ n **information** et α **risque** donnés \rightarrow quelle **précision** ε ?
 - ▶ risque α et précision ε donnés \rightarrow quel nombre minimal de données n nécessaires ?
 - ▶ quel risque prend-on si l'on suppose une précision ε avec n données ?
- ▶ Plusieurs approches :
 - ▶ non-asymptotique naïve
 - ▶ non-asymptotique
 - ▶ **approche asymptotique** (via des théorèmes limites)

Inégalité de Markov

- ▶ Si Y est une v.a. positive et $t \geq 0$, $Y \mathbb{1}_{\{Y \geq t\}} \geq t \mathbb{1}_{\{Y \geq t\}}$
- ▶ Inégalité de Markov

$$\mathbb{P}(Y \geq t) \leq t^{-1} \mathbb{E}[Y].$$

- ▶ Si ϕ est une fonction positive monotone croissante, $\phi(t) > 0$ pour tout $t > 0$,

$$\mathbb{P}(Y \geq t) = \mathbb{P}(\phi(Y) \geq \phi(t)) \leq \mathbb{E}[\phi(Y)] / \phi(t).$$

- ▶ Bien entendu, cette inégalité est intéressante ssi $\mathbb{E}[\phi(Y)] < \infty$.

Approche naïve : contrôle de la variance

Soit $\alpha > 0$ donné (petit). On veut trouver ε , le plus petit possible, de sorte que

$$\mathbb{P} \left(|\hat{F}_n(x_0) - F(x_0)| \geq \varepsilon \right) \leq \alpha.$$

On a (Tchebychev)

$$\begin{aligned} \mathbb{P} [|\hat{F}_n(x_0) - F(x_0)| \geq \varepsilon] &\leq \frac{1}{\varepsilon^2} \text{Var}[\hat{F}_n(x_0)] \\ &= \frac{F(x_0)(1 - F(x_0))}{n\varepsilon^2} \\ &\leq \frac{1}{4n\varepsilon^2} \\ &\leq \alpha \end{aligned}$$

Conduit à

$$\varepsilon = \frac{1}{2\sqrt{n\alpha}}$$

Intervalle de confiance

Conclusion : pour tout $\alpha > 0$,

$$\mathbb{P} \left[|\hat{F}_n(x_0) - F(x_0)| \geq \frac{1}{2\sqrt{n\alpha}} \right] \leq \alpha.$$

Terminologie

L'intervalle

$$\mathcal{I}_{n,\alpha} = \left[\hat{F}_n(x_0) \pm \frac{1}{2\sqrt{n\alpha}} \right]$$

est un intervalle de confiance pour $F(x_0)$ au niveau de confiance $1 - \alpha$.

Précision catastrophique !

- ▶ Si $\alpha = 5\%$ et $n = 100$, précision $\varepsilon = 0.22$.
- ▶ Autres exemples : $\varepsilon_{\alpha=1/1000, n=100} = 1.58$,
 $\varepsilon_{\alpha=1/100, n=100} = 0.5$. aucune précision d'estimation !
- ▶ D'où vient le défaut de cette précision ?
 - ▶ Mauvais choix de l'estimateur ? (\rightarrow on verra que **non**).
 - ▶ Mauvaise estimation de l'erreur ?

Inégalité de Markov

- **Moments d'ordres plus élevés** : Pour tout $q > 0$, on a en posant $\phi(t) = t^q$

$$\mathbb{P}(|\hat{F}_n(x_0) - F(x_0)| \geq \varepsilon) \leq \frac{1}{\varepsilon^q} \mathbb{E}(|F_n(x_0) - F(x_0)|^q)$$

- ... ‡ part pour $q = 2$ (ou plus généralement q entier pair), $\mathbb{E}(|F_n(x_0) - F(x_0)|^q)$ ne se calcule pas très aisément...
- Plus intéressant de considérer une inégalité **exponentielle**.

Inégalité exponentielle

- ▶ On pose $\phi(t) = e^{\lambda t}$. Dans ce cas, l'inégalité de Markov implique

$$\mathbb{P}(Z > t) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda Z}]$$

où $\Phi(\lambda) = \mathbb{E}[e^{\lambda Z}]$ est la **fonction génératrice** des moments ou *transformée de Laplace*.

- ▶ En notant $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$ le logarithme la transformée de Laplace et en introduisant

$$\psi_Z^*(t) = \sup_{\lambda \geq 0} \{ \lambda t - \psi_Z(\lambda) \}$$

nous obtenons la borne de **Cramér-Chernoff**

$$\mathbb{P}(Z > t) \leq \exp(-\psi_Z^*(t)).$$

Inégalité de Hoeffding

Proposition

Y_1, \dots, Y_n i.i.d. de loi de Bernoulli de paramètre p . Alors

$$\mathbb{P} \left(\left| n^{-1} \sum_{i=1}^n Y_i - p \right| \geq t \right) \leq 2 \exp(-2nt^2).$$

Application : on pose $Y_i = 1_{\{x_i \leq x_0\}}$ et $p = F(x_0)$. On en déduit

$$\mathbb{P} [|\hat{F}_n(x_0) - F(x_0)| \geq \varepsilon] \leq 2 \exp(-2n\varepsilon^2).$$

On résout en ε :

$$2 \exp(-2n\varepsilon^2) = \alpha,$$

soit

$$\varepsilon = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}.$$

Comparaison Tchebychev vs. Hoeffding

Nouvel intervalle de confiance

$$\mathcal{I}_{n,\alpha}^{\text{hoeffding}} = \left[\hat{F}_n(x_0) \pm \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \right],$$

à comparer avec

$$\mathcal{I}_{n,\alpha}^{\text{tchebychev}} = \left[\hat{F}_n(x_0) \pm \frac{1}{2\sqrt{n\alpha}} \right].$$

- ▶ Même ordre de grandeur en n .
- ▶ Gain **significatif** dans la limite $\alpha \rightarrow 0$. La « prise de risque » devient marginale par rapport au nombre d'observations.
- ▶ **Optimalité d'une telle approche ?**

L'approche asymptotique

- Vers une notion d'optimalité : on se place dans la limite $n \rightarrow \infty$ (l'information « explose »). On évalue

$$\mathbb{P} \left[\left| \hat{F}_n(x_0) - F(x_0) \right| \geq \varepsilon \right], n \rightarrow \infty$$

pour une normalisation $\varepsilon = \varepsilon_n$ appropriée.

- Outil : **Théorème central-limite.**

Convergence en loi

La suite $(X_n)_{n \geq 0}$ converge en loi vers X ($X_n \xrightarrow{d} X$) ssi l'une des conditions **équivalente** est vérifiée :

- Pour toute fonction f continue bornée,

$$\lim_{n \rightarrow \infty} \mathbb{E}[f(X_n)]$$



$$\mathbb{P}(X_n \leq x) \rightarrow \mathbb{P}(X \leq x)$$

en tout point x où la fonction de répartition de X est continue

- Pour tout $u \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{E}[e^{iuX_n}] = \mathbb{E}[e^{iuX}].$$

Convergence en loi

- **Attention...** ce sont les lois qui **convergent**... Si X et $-X$ ont la même loi (par exemple, $X \sim \mathcal{N}(0, 1)$), on a simultanément

$$X_n \xrightarrow{d} X \quad \text{et} \quad X_n \xrightarrow{d} -X$$

- On peut avoir $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{d} Y$ **sans avoir** $(X_n, Y_n) \xrightarrow{d} (X, Y)$ (on n'a d'ailleurs pas spécifié la loi jointe du couple (X, Y))
- Par contre, si $(X_n, Y_n) \xrightarrow{d} (X, Y)$, on a pour toute fonction ϕ continue, $\phi(X_n, Y_n) \xrightarrow{d} \phi(X, Y)$, et donc $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{d} Y$.

Rappel : théorème central-limite

Theoreme

Si Y_1, \dots, Y_n i.i.d., $\mu = \mathbb{E}[Y_i]$, $0 < \sigma^2 = \text{Var}[Y_i] < +\infty$, alors

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Y_i - \mu \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Interprétation et application

- Interprétation du TCL :

$$\frac{1}{n} \sum_{i=1}^n Y_i = \mu + \frac{\sigma}{\sqrt{n}} \xi^{(n)}, \quad \xi^{(n)} \overset{d}{\approx} \mathcal{N}(0, 1).$$

- Application : $Y_i = 1_{\{X_i \leq x_0\}}$, $\mu = F(x_0)$,
 $\sigma(\textcolor{red}{F}) = F(x_0)^{1/2}(1 - F(x_0))^{1/2}$.

On a

$$\begin{aligned} \mathbb{P} \left[\left| \hat{F}_n(x_0) - F(x_0) \right| \geq \varepsilon_n \right] &= \mathbb{P} \left[\left| \xi^{(n)} \right| \geq \frac{\sqrt{n} \varepsilon_n}{\sigma(\textcolor{red}{F})} \right] \\ &= \mathbb{P} \left[\left| \xi^{(n)} \right| \geq \frac{\varepsilon_0}{\sigma(\textcolor{red}{F})} \right] \end{aligned}$$

pour la calibration $\varepsilon_n = \varepsilon_0 / \sqrt{n}$ (ε_0 reste à choisir).

TCL et intervalle de confiance (suite)

Il vient

$$\begin{aligned}\mathbb{P} \left[\left| \xi^{(n)} \right| \geq \frac{\varepsilon_0}{\sigma(\textcolor{red}{F})} \right] &\rightarrow \int_{|x| \geq \varepsilon_0/\sigma(\textcolor{red}{F})} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} \\ &= 2 \left(1 - \Phi(\varepsilon_0/\sigma(\textcolor{red}{F})) \right) \\ &\leq \alpha,\end{aligned}$$

avec $\Phi(x) = \int_{-\infty}^x e^{-t^2/2} dt$, ce qui donne

$$\boxed{\varepsilon_0 = \sigma(\textcolor{red}{F}) \Phi^{-1}(1 - \alpha/2).}$$

TCL et intervalle de confiance : (suite)

- ▶ On a montré

$$\mathbb{P} \left[\left| \hat{F}_n(x_0) - F(x_0) \right| \geq \frac{\sigma(F)}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) \right] \rightarrow \alpha.$$

- ▶ Attention ! ceci ne fournit **pas** un intervalle de confiance :
 $\sigma(F) = F(x_0)^{1/2} (1 - F(x_0))^{1/2}$ est inconnu !
- ▶ Solution : remplacer $\sigma(F)$ par $\hat{F}_n(x_0)^{1/2} (1 - \hat{F}_n(x_0))^{1/2}$ observable.

Lemme de Slutsky

Lemme

Si $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{\mathbb{P}} c$ (constante), alors $(X_n, Y_n) \xrightarrow{d} (X, Y)$.

TCL et intervalle de confiance : conclusion

Proposition

Pour tout $\alpha \in (0, 1)$,

$$\mathcal{I}_{n,\alpha}^{\text{asympt}} = \left[\hat{F}_n(x_0) \pm \frac{\hat{F}_n(x_0)^{1/2} (1 - \hat{F}_n(x_0))^{1/2}}{\sqrt{n}} \Phi^{-1}(1 - \alpha/2) \right]$$

est un intervalle de confiance asymptotique pour $F(x_0)$ au niveau de confiance $1 - \alpha$:

$$\mathbb{P} \left[F(x_0) \in \mathcal{I}_{n,\alpha}^{\text{asympt}} \right] \rightarrow 1 - \alpha.$$

Le passage $\sigma(\textcolor{red}{F}) \longrightarrow \hat{F}_n(x_0)^{1/2} (1 - \hat{F}_n(x_0))^{1/2}$ est licite via le lemme de Slutsky.

