

Statistiques mathématiques : cours 1

Guillaume Lécué

2 septembre 2016

Organisation

9 Cours (18h) Guillaume Lécué

guillaume.lecue@ensae.fr

Les **mercredis** de 11h à 13h et **vendredis** de septembre de 14h à 16h (sauf le vendredi 30) et **mercredi 5 octobre**.

Slides du cours téléchargeable à

<http://lecueguillaume.github.io/2015/10/05/rappels-stats/>

6 TD (12h) Philippe Mesnard

Les **mercredis** de septembre et le **mercredi 5 octobre** de 14h à 16h et **lundi 19 septembre** de 11h à 13h.

Examen

Fin octobre/ début novembre

Présentation (succinte) du cours de stats math

- ▶ Echantillonnage et modélisation statistique. Fonction de répartition empirique (2 cours)
- ▶ Méthodes d'estimation classiques (2 cours)
- ▶ Information statistique, théorie asymptotique pour l'estimation (2 cours)
- ▶ Décision statistique et tests (2 cours)
- ▶ Compléments sur le modèle linéaire et statistiques Bayésiennes(1 cours)

Aujourd'hui

Organisation du cours

Echantillonnage et modélisation statistique

- Données d'aujourd'hui

- Expérience statistique

- Modèle statistique

Fonction de répartition empirique et théorème fondamentale de la statistique

- Loi d'une variable aléatoire

- Fonction de répartition empirique

- Approche non-asymptotique

Les données d'aujourd'hui : fichiers (en local) .csv ou .txt

Les chiffres du travail

Taux d'activité par tranche d'âge hommes vs. femmes

	A	B	C	D	E	F	G	H	I
1									
2	Taux d'activité par tranche d'âge de 1975 à 2005								
3	En %								
4		1975	1976	1977	1978	1979	1980	1981	1982
5	Femmes								
6	15-24 ans	45,5	45,7	45,2	43,9	44,2	42,9	42,1	41,87
7	25-49 ans	58,6	60,3	62,1	62,8	64,7	65,4	66,2	67,55
8	50 ans et plus	42,9	43,1	44,4	43,9	44,8	45,9	45,2	43,47
9	Ensemble	51,5	52,5	53,6	53,6	54,8	55,1	55,1	55,29
10	Hommes								
11	15-24 ans	55,6	54,7	53,7	52,2	52,5	52,0	50,4	45,02
12	25-49 ans	97,0	97,1	96,9	96,9	96,9	97,1	96,9	96,75
13	50 ans et plus	79,5	78,8	79,5	78,8	79,4	78,3	75,4	71,65
14	Ensemble	82,5	82,2	82,1	81,6	81,8	81,5	80,4	78,14

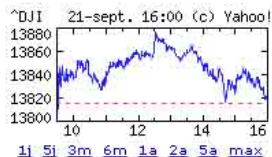
<http://www.insee.fr/>
<https://www.data.gouv.fr/>

Les données d'aujourd'hui : séries temporelles

Le monde de la finance

DOW JONES INDUSTRIAL AVERAGE IN (DJI: ^DJI)

Dern. Cours:	13.820,19
Heure:	21 sept.
Variation:	↑ 53,49 (0,39%)
Clôture Préc.:	13.766,70
Ouverture:	13.768,33
Var. Journalière:	13.768,25 - 13.877,17
Var. sur 1 an:	11.926,80 - 14.121,00
Volume:	419.389.397

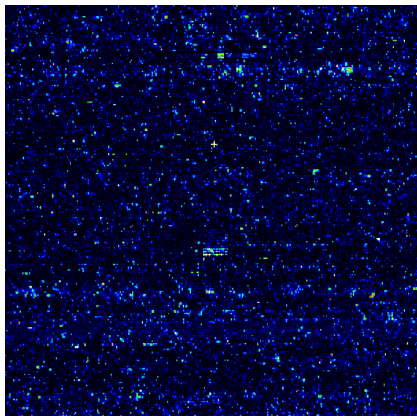


<http://fr.finance.yahoo.com/>

<http://www.bloomberg.com/enterprise/data/>

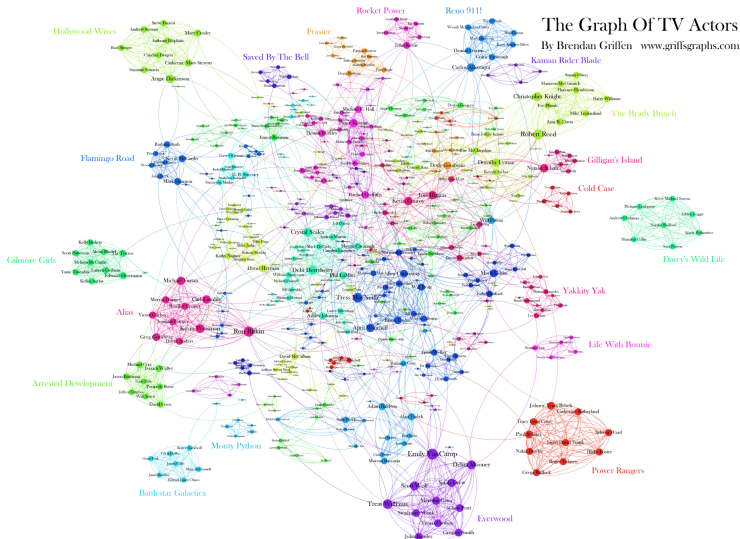
Les données d'aujourd'hui : grandes matrices

Biopuces et analyse d'ADN



Les données d'aujourd'hui : graphes

acteurs de séries



Les données d'aujourd'hui : le métier en data science

Problématique :

- ▶ stockage, requettage : expertise en base de données
- ▶ data “jujitsu”, data “massage”
- ▶ data-vizualization (Gephi, Tulip, widget python, etc.)
- ▶ mathématiques :
 - ★ **modélisation** (statistiques)
 - ★ **construction d'estimateurs**
implémentation d'algorithmes
- ▶ Python, R, H2O, TensorFlow, vowpal wabbit, spark,..., github,...

Pour s'entrainer aux métiers en “data science” :

- <https://www.kaggle.com>, <https://www.datascience.net/>
- notebooks python
- Coursera

Objectif du cours “statistiques mathématiques”

1. Construire des modèles statistiques pour des données classiques
2. Construire des estimateurs / tests classiques
3. Connaître leurs propriétés statistiques et les outils mathématiques qui permettent de les obtenir

Problématique statistique

- 1) **Point de départ** : données (ex. : des nombres réels)

$$\mathbf{x}_1, \dots, \mathbf{x}_n$$

- 2) **Modélisation statistique** :

- ▶ les données sont des réalisations

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

(autrement dit, pour un certain ω , $X_1(\omega) = \mathbf{x}_1, \dots, X_n(\omega) = \mathbf{x}_n$)

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

$$\boxed{\{\mathbb{P}_\theta^n, \theta \in \Theta\}} : \text{le modèle}$$

On pense qu'il existe $\theta \in \Theta$ tel que $\mathbb{P}^{(X_1, \dots, X_n)} = \mathbb{P}_\theta^n$.

- 3) **Problématique** : à partir de « l'observation » X_1, \dots, X_n , peut-on **estimer** θ ? **tester** des propriétés de 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 des données X_1, \dots, X_n , établir une **décision** $\varphi_n(X_1, \dots, X_n) \in \{\text{ensemble de décisions}\}$ concernant une hypothèse sur θ .

Definition

Une **statistique** est une fonction mesurable des données

!ATTENTION! Une statistique ne peut pas dépendre du paramètre inconnu : une statistique se construit uniquement à partir des données !

Exemple du pile ou face

- ▶ 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** $\theta \in \Theta = [0, 1]$.
 - ▶ **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 $|\bar{X}_{18} - 0.5| \ll \text{petit}$, on accepte l'hypothèse « la pièce est équilibrée ». Sinon, on rejette. Quel seuil choisir, et avec quelles conséquences (ex. probabilité de se tromper).

Echantillonnage = répétition d'une même expérience

- ▶ 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 (i.i.d.)**, de même loi commune $\mathbb{P}^X \in \{\mathbb{P}_\theta : \theta \in \Theta\}$.
- ▶ problème : à partir des données X_1, \dots, X_n que dire de la loi \mathbb{P}^X commune aux X_i ? (moyenne, moments, symétrie, densité, etc.)

Expérience statistique

Consiste à déterminer :

- ▶ l'espace des observations

$$\mathfrak{Z} \text{ (ex. : } \mathfrak{Z} = \{0, 1\}^{18}\text{)}$$

C'est l'espace où vivent les observations

- ▶ Une **tribu** : \mathcal{Z} (on modélise les données comme des réalisations de variables aléatoires...) (ex. : $\mathcal{Z} = \mathcal{P}(\mathfrak{Z})$)
- ▶ Une famille de lois = modèle

$$\{\mathbb{P}_\theta, \theta \in \Theta\} \text{ (ex. : } \mathbb{P}_\theta = \mathbb{P}_\theta^n = (\theta\delta_1 + (1 - \theta)\delta_0)^{\otimes 18}\text{)}$$

Expérience statistique

Definition

Une *expérience statistique* \mathcal{E} est un triplet

$$\mathcal{E} = (\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$$

où

- ▶ $(\mathfrak{Z}, \mathcal{Z})$ espace mesurable (ex. : $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$),
- ▶ $\{\mathbb{P}_\theta, \theta \in \Theta\}$ famille de probabilités définies *simultanément* sur le même espace $(\mathfrak{Z}, \mathcal{Z})$.

Modèles statistiques (jargon)

- ▶ $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est appelé **modèle**
- ▶ quand il existe k tel que $\Theta \subset \mathbb{R}^k$, on parle de modèle **paramétrique**
- ▶ quand θ est un paramètre infini dimensionnel, on parle de modèle **non-paramétrique** (ex. : densité)
- ▶ quand $\theta = (f, \theta_0)$ où f est infini dimensionnel (souvent, paramètre de nuisance) et $\theta_0 \in \mathbb{R}^k$ (paramètre d'intérêt), on parle de modèle **semi-paramétrique**
- ▶ quand $\theta \in \Theta \mapsto \mathbb{P}_\theta$ est injectif, on dit que le modèle est **identifiable**

Question centrale en statistiques : Quel modèle est le plus adapté à ces données ?

Il existe deux manières équivalentes de définir un modèle :

1. soit en se donnant une famille de loi $\{\mathbb{P}_\theta, \theta \in \Theta\}$
2. soit en se donnant une équation

Exemple de modèle/modélisation (1)

On observe un n -uplet de variables aléatoires réelles :

$$Z = (X_1, \dots, X_n)$$

On peut modéliser ces observations de deux manières (équivalentes) :

- Famille de lois : $\{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$, par exemple,

$$\mathbb{P}_\theta = (\mathcal{N}(\theta, 1))^{\otimes n}$$

- Par une équation : pour tout $i \in 1, \dots, n$,

$$X_i = \theta + g_i$$

où g_1, \dots, g_n sont n variables aléatoires Gaussiennes centrées réduites indépendantes.

Exemple de modèle/modélisation (2)

On observe un n -uplet de variables aléatoires réelles :

$$Z = (X_1, \dots, X_n)$$

On peut modéliser ces observations de deux manières (équivalentes) :

- ▶ Par une équation : $X_1 = g_1$ et pour tout $i \in 1, \dots, n-1$,

$$X_{i+1} = \theta X_i + g_i$$

où g_1, \dots, g_n sont iid $\mathcal{N}(0, 1)$.

- ▶ Famille de lois : $\{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ où

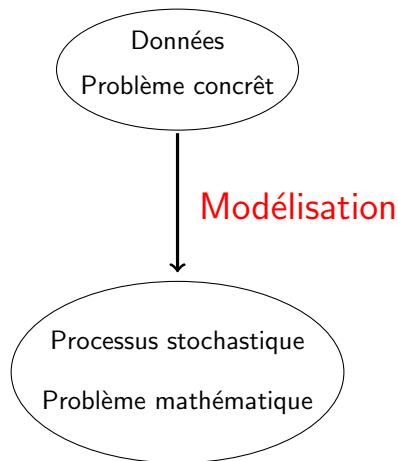
$$\mathbb{P}_\theta = f_\theta \cdot \lambda^n$$

où λ^n est la mesure de Lebesgue sur \mathbb{R}^n et

$$f_\theta(x_1, \dots, x_n) = f(x_1)f(x_2 - \theta x_1) \cdots f(x_n - \theta x_{n-1})$$

$$\text{et } f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}}.$$

Pourquoi modéliser ?



Pourquoi modéliser ? :

- 1) Outils mathématiques
- 2) Résultats mathématiques
- 3) Algorithmes

3 modèles (non-paramétriques) classiques

1. Modèle de **densité** : on observe un n -échantillon

X_1, \dots, X_n de v.a.r. de densité f tel que $f \in \mathcal{C}$

où \mathcal{C} est une classe de densités sur \mathbb{R} (Lebesgue).

2. Modèle de **régression** : on observe un n -échantillon de couples $(X_i, Y_i)_{i=1}^n$ tel que $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^d$ et

$$Y_i = f(X_i) + \xi_i$$

où ξ_i sont des v.a.r.i.i.d. indépendantes des X_i et $f \in \mathcal{C}$.

- ▶ quand $f(X_i) = \langle \theta, X_i \rangle$: modèle de regression **linéaire**,
- ▶ et quand $\xi_i \sim \mathcal{N}(0, \sigma^2)$: modèle **linéaire Gaussien**

3. modèle de **classification** : on observe un n -échantillon $(Y_i, X_i)_{i=1}^n$ tel que $Y_i \in \{0, 1\}$ et $X_i \in \mathcal{X}$. Par ex. :

$$\mathbb{P}[Y_i = 1 | X_i = x] = \sigma(\langle x, \theta \rangle) \text{ où } \sigma(x) = (1 + e^{-x})^{-1}$$

Fonction de répartition empirique et théorème fondamentale de la statistique

Question fondamentale

Considérons le modèle d'échantillonnage sur \mathbb{R} : on observe

$$X_1, \dots, X_n$$

qui sont i.i.d. de loi commune \mathbb{P}_X .

Rem. : Comme la loi de l'observation (X_1, \dots, X_n) est $\mathbb{P}_X^{\otimes n}$, se donner un modèle est ici (pour le modèle d'échantillonnage) équivalent à se donner un modèle sur \mathbb{P}_X .

Par exemple : $\mathbb{P}_X \in \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$

Question fondamentale

On considère le modèle "total" = $\mathbb{P}_X \in \{\text{toutes les lois sur } \mathbb{R}\}$, est-il possible de connaître **exactement** \mathbb{P}_X quand le nombre n de données tends vers ∞ ?

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

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

Formule d'intégration

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

pour toute fonction test φ .

Loi d'une variable aléatoire (1/4)

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 :

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

- Formule de calcul (φ fonction test)

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

Loi d'une variable aléatoire (2/4)

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

- ▶ La loi de X est décrite par

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

- ▶ Ecriture de \mathbb{P}^X :

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

- ▶ Formule de calcul (φ fonction test)

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

Loi d'une variable aléatoire (3/4)

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 :

$$\boxed{\mathbb{P}^X = f \cdot \lambda} \text{ où } f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

λ : 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 (4/4)

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

- Ecriture de \mathbb{P}^X :

$$\mathbb{P}^X = g \cdot \lambda + \mathbb{P}[Z \geq 1] \delta_1,$$

où $g(x) = f(x)I(x < 1), \forall x \in \mathbb{R}$.

- Formule de calcul

$$\mathbb{E}[\varphi(X)] = \int_{-\infty}^1 \varphi(x) f(x) dx + \mathbb{P}[Z \geq 1] \varphi(1)$$

Fonction de répartition

Les lois sont des objets compliquées. On peut néanmoins les caractériser par des objets plus simples.

Definition

Soit X variable aléatoire réelle. La fonction de répartition de X est :

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

- ▶ 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** de X désignera indifféremment F ou \mathbb{P}^X .

Retour sur la question fondamentale

On « observe »

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

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

Question : Est-il possible de retrouver exactement F quand n tends vers ∞ ?

Idée : On va chercher à estimer F sur \mathbb{R} . Soit $x \in \mathbb{R}$. $F(x) = \mathbb{P}[X \leq x]$ est la probability que X soit plus petit que x . On va alors compter le nombres de X_i qui sont plus petit que x et diviser par n :

$$\frac{1}{n} \sum_{i=1}^n I(X_i \leq x).$$

Fonction de répartition empirique

Definition

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

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

(C'est une fonction aléatoire)

Propriétés asymptotiques de $\hat{F}_n(x)$

Pour tout $x \in \mathbb{R}$:

$$\hat{F}_n(x) \xrightarrow{p.s.} F(x) \text{ quand } n \rightarrow \infty$$

C'est une conséquence de la **loi forte des grands nombres** appliquée à la suite de v.a.r.i.i.d. $(I(X_i \leq x))_i$.

On dit que $\hat{F}_n(x)$ est un estimateur **fortement consistant** de $F(x)$.

Propriétés asymptotiques de \hat{F}_n

Theorem (Glivenko-Cantelli)

$$\left\| \hat{F}_n - F \right\|_{\infty} \xrightarrow{p.s.} 0 \text{ quand } n \rightarrow \infty$$

Aussi appelé **Théorème fondamental de la statistique**.

Interprétation : Avec un nombre infini de données dans le modèle d'échantillonnage, on peut donc reconstruire exactement F et donc déterminer exactement la loi des observations.

Notebooks

`http://localhost:8888/notebooks/cdf_empirique.ipynb`
Glivenko-Cantelli

Autres propriétés asymptotiques de $\hat{F}_n(x)$

Soit $x \in \mathbb{R}$. On sait que si $n \rightarrow \infty$ alors

$$\hat{F}_n(x) \xrightarrow{p.s.} F(x)$$

Question : Quelle est la vitesse de convergence de $F_n(x)$ vers $F(x)$?

Outil : **Théorème central-limite** appliqué à la suite de v.a.r.i.i.d.

$(I(X_i \leq x))_i$:

$$\sqrt{n}(\hat{F}_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1 - F(x)))$$

On dit que $\hat{F}_n(x)$ est **asymptotiquement normal** de **variance asymptotique** $F(x)(1 - F(x))$.

Notebooks

`http://localhost:8888/notebooks/cdf_empirique.ipynb`
Glivenko-Cantelli

TCL et intervalle de confiance asymptotique

On a montré par le TCL que pour tout $0 < \alpha < 1$, quand $n \rightarrow \infty$,

$$\mathbb{P} \left[|\hat{F}_n(x) - F(x)| \geq c_\alpha \frac{\sigma(F)}{\sqrt{n}} \right] \rightarrow \int_{|x| > c_\alpha} \exp(-x^2/2) \frac{dx}{\sqrt{2\pi}} = \alpha$$

où $\sigma(F) = F(x)(1 - F(x))$ et $c_\alpha = \Phi^{-1}(1 - \alpha/2)$.

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

TCL et intervalle de confiance asymptotique

Proposition

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

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

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

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

Theorem (Théorème de Kolmogorov-Smirnov)

Soit X une v.a.r. de fonction de répartition F qu'on suppose continue et $(X_n)_n$ une suite de v.a.r. i.i.d. de même loi que X alors :

$$\sqrt{n} \left\| \hat{F}_n - F \right\|_{\infty} \xrightarrow{d} K$$

où K est une variable aléatoire telle que pour tout $x \in \mathbb{R}$

$$\mathbb{P}[K \leq x] = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 x^2)$$

- ▶ Utile pour le **test de Kolmogorov-Smirnov**
- ▶ version non-asymptotique de ce résultat : quand F est continue, la loi de $\left\| \hat{F}_n - F \right\|_{\infty}$ est indépendante de F

résultats asymptotiques et non-asymptotiques

On classe les résultats statistiques en deux catégories :

1. Un résultat obtenu quand n tend vers l'infini est un résultat dit asymptotique
2. Un résultat obtenu à n fixé est un résultat dit non-asymptotique

Estimation non-asymptotique de $F(x)$ par $\hat{F}_n(x)$

Soit $0 < \alpha < 1$ donné (petit). On veut trouver ε , le plus petit possible, de sorte que

$$\mathbb{P} [|\hat{F}_n(x) - F(x)| \geq \varepsilon] \leq \alpha.$$

On a (Tchebychev)

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

Conduit à

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

Intervalle de confiance

Conclusion : pour tout $\alpha > 0$,

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

Terminologie

L'intervalle

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

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

Inégalité de Hoeffding

Proposition

Y_1, \dots, Y_n v.a.r.i.i.d. telles que $a \leq Y_1 \leq b$ p.s.. Alors

$$\mathbb{P} \left[\left| \frac{1}{n} \sum_{i=1}^n Y_i - \mathbb{E} Y_1 \right| \geq t \right] \leq 2 \exp \left(- \frac{2nt^2}{(a-b)^2} \right)$$

Application : on fait $Y_i = I(x_i \leq x)$ et $p = F(x)$. On en déduit

$$\mathbb{P} \left[\left| \widehat{F}_n(x) - F(x) \right| \geq \varepsilon \right] \leq 2 \exp(-2n\varepsilon^2).$$

On résout en ε :

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

soit

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

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 ?**

Observation finale

Comparaison des longueurs des 3 intervalles de confiance :

- ▶ Tchebychev (non-asymptotique) $\frac{2}{\sqrt{n}} \frac{1}{2} \frac{1}{\sqrt{\alpha}}$
- ▶ Hoeffding (non-asymptotique) $\frac{2}{\sqrt{n}} \sqrt{\frac{1}{2} \log \frac{2}{\alpha}}$
- ▶ TCL (asymptotique) $\frac{2}{\sqrt{n}} \hat{F}_n(x_0)^{1/2} (1 - \hat{F}_n(x_0))^{1/2} \Phi^{-1}(1 - \alpha/2)$.
- ▶ La longueur la plus petite est (**sans surprise !**) celle fournie par le TCL. Mais la longueur de l'intervalle de confiance fournie par l'inégalité de Hoeffding **comparable** au TCL en n et α (dans la limite $\alpha \rightarrow 0$).

Version non-asymptotique de Kolmogorov-Smirnov

X_1, \dots, X_n i.i.d. de loi F continue, \hat{F}_n leur fonction de répartition empirique.

Proposition (Inégalité de Dvoretzky-Kiefer-Wolfowitz)

Pour tout $\varepsilon > 0$.

$$\mathbb{P} \left[\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \geq \varepsilon \right] \leq 2 \exp(-2n\varepsilon^2).$$

- ▶ Résultat difficile (théorie des processus empiriques).
- ▶ Permet de construire des régions de confiance avec des résultats similaires au cadre ponctuel :

$$\mathbb{P} \left[\forall x \in \mathbb{R}, F(x) \in [\hat{F}_n(x) \pm \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}}] \right] \geq 1 - \alpha$$

Rappels de probabilités

Tribus et mesures de probabilité

Soit \mathfrak{J} un ensemble.

1. Une **tribu** \mathcal{Z} sur \mathfrak{J} est un ensemble de parties de \mathfrak{J} tel que :
 - ▶ \mathcal{Z} est stable par union et intersection dénombrable
 - ▶ \mathcal{Z} est stable par passage au complémentaire
 - ▶ $\mathfrak{J} \in \mathcal{Z}$

Les éléments de \mathcal{Z} sont appelés des **événements**.

2. Une **mesure de probabilité** sur $(\mathfrak{J}, \mathcal{Z})$ est une application $\mathbb{P} : \mathcal{Z} \mapsto [0, 1]$ telle que
 - ▶ $\mathbb{P}[\mathfrak{J}] = 1$
 - ▶ Si (A_n) est une famille dénombrable d'événements disjoints alors

$$\mathbb{P} \left[\bigcup_n A_n \right] = \sum_n \mathbb{P}[A_n]$$

Le dernier point est aussi équivalent à : pour (A_n) une suite croissante d'événements on a $\mathbb{P}(A_n) \uparrow \mathbb{P}(\bigcup A_n)$.

Type de convergence de suite de variables aléatoires

Soit (Z_n) une suite de variable aléatoires et Z une variable aléatoire à valeurs dans (\mathbb{R}, \mathbb{B}) (toutes définies sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$).

1. (Z_n) converge en **loi** vers Z , noté $Z_n \xrightarrow{d} Z$, quand pour toute fonction continue bornée $f : \mathbb{R} \mapsto \mathbb{R}$ on a

$$\mathbb{E} f(Z_n) \rightarrow \mathbb{E} f(Z)$$

2. (Z_n) converge en **probabilité**, vers Z , noté $Z_n \xrightarrow{\mathbb{P}} Z$, quand pour tout $\epsilon > 0$,

$$\mathbb{P} [|Z_n - Z| \geq \epsilon] \rightarrow 0$$

3. (Z_n) converge **presque sûrement** vers Z , noté $Z_n \xrightarrow{p.s.} Z$, quand il existe un événement $\Omega_0 \in \mathcal{F}$ tel que $\mathbb{P}[\Omega_0] = 1$ et pour tout $\omega \in \Omega_0$

$$Z_n(\omega) \rightarrow Z(\omega)$$

Loi forte des grands nombres

Theorem

Soit (X_n) une suite de v.a.r.i.i.d. telle que $\mathbb{E}|X_1| < \infty$. Alors

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p.s.} \mathbb{E} X_1$$

Il y a aussi une “équivalence” à ce résultat : si (X_n) est une suite de v.a.r.i.i.d. telle que $\left(\frac{1}{n} \sum_{i=1}^n X_i\right)_n$ converge presque sûrement alors $\mathbb{E}|X_1| < \infty$ et elle converge presque sûrement vers $\mathbb{E} X_1$.

Théorème central-limite

Theorem

Soit (X_n) une suite de v.a.r.i.i.d. telle que $\mathbb{E} X_1^2 < \infty$. Alors

$$\frac{\sqrt{n}}{\sigma} \left(\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E} X_1 \right) \xrightarrow{d} \mathcal{N}(0, 1)$$

- ▶ TCL : « vitesse » dans la loi des grands nombres.
- ▶ Interprétation du TCL :

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

- ▶ Le mode de convergence est **la convergence en loi**. Ne peut pas avoir lieu en probabilité.

Lemme de Slutsky

- ▶ Le vecteur $(X_n, Y_n) \xrightarrow{d} (X, Y)$ si

$$\mathbb{E} [\varphi(X_n, Y_n)] \rightarrow \mathbb{E} [\varphi(X, Y)],$$

pour φ continue bornée.

- ▶ **Attention !** Si $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{d} Y$, on n'a pas en général $(X_n, Y_n) \xrightarrow{d} (X, Y)$.
- ▶ **Mais** (lemme de Slutsky) si $X_n \xrightarrow{d} X$ et $Y_n \xrightarrow{\mathbb{P}} c$ (constante), alors $(X_n, Y_n) \xrightarrow{d} (X, Y)$.
- ▶ Par suite, sous les hypothèses du lemme, pour toute fonction continue g , on a $g(X_n, Y_n) \xrightarrow{d} g(X, Y)$.

Continuous map theorem

Soit $f : \mathbb{R} \mapsto \mathbb{R}$ une fonction continue et (X_n) une suite de v.a.r.

1. si (X_n) converge en **loi** vers X alors $f(X_n)$ converge en loi vers $f(X)$
2. si (X_n) converge en **probabilité** vers X alors $f(X_n)$ converge en probabilité vers $f(X)$
3. si (X_n) converge **p.s.** vers X alors $f(X_n)$ converge p.s. vers $f(X)$