Rappels de statistiques mathématiques : cours 1

Guillaume Lecué

2 septembre 2015

Aujourd'hui

Organisation

Cours Guillaume Lecué

guillaume.lecue@cmap.polytechnique.fr

Les lundis et mercredis de septembre de 11h à 13h

TD Vincent Cottet

Les lundis de septembre et les vendredis 21 et 28 septembre de 14h à 16h.

Examen

Fin octobre/ début novembre

Présentation (succinte) du cours de rappels stats

- 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 modéle linéaire et statistiques Bayésiennes(1 cours)

Plan

- Données et modélisation statistique
- fonction de répartition empirique

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

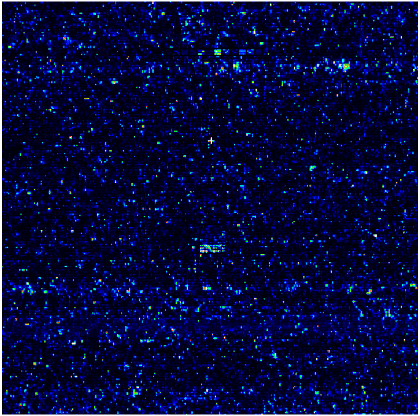
DOW JONES INDUSTRIAL AVERAGE IN (DJI:^DJI)

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/

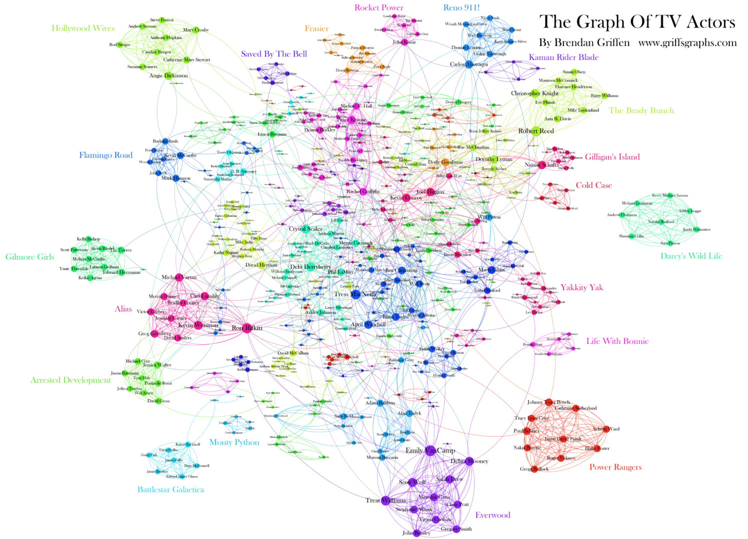
Les données d'aujourd'hui : grandes matrices

Biopuces et analyse d'ADN



Les données d'aujourd'hui : graphes

acteurs de séries



Problèmatique :

- 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, Torch7, 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 "Rappels de 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)

$$\mathtt{x_1, \dots, x_n}$$

2) Modélisation statistique :

- les données sont des réalisations

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

- 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)

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

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 $X_1, \ldots, X_n$, construire $\varphi_n(X_1, \ldots, X_n)$ qui « approche au mieux » θ.
- Test : à partir des données $X_1, \ldots, X_n$, établir une décision $\varphi_n(X_1, \ldots, X_n) \in \{$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 $(P = 0,\ F = 1)$

$$0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0$$

- Modéle statistique : on observe $n = 18$ variables aléatoires X_i indépendantes, de Bernoulli de paramètre inconnu $\theta \in \Theta = [0, 1]$.
  - Estimation. Estimateur $\bar{X}_{18} = \frac{1}{18}\sum_{i=1}^{18} X_i \overset{\text{ici}}{=} 8/18 = 0.44$. 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 $=$ 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, 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

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

  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 = modéle

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

Expérience statistique

Definition

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

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

où

- $\left(\mathfrak{Z}, \mathcal{Z}\right)$ *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* $\left(\mathfrak{Z}, \mathcal{Z}\right)$.

Modéles statistiques (jargon)

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

Modéles statistiques

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

Il existe deux manières équivalentes de définir un modéle :

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

Exemple de modéle/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,

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

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

$$\boxed{X_i = \theta + g_i}$$

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

Exemple de modéle/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 $i \in 1, \dots, n-1$,

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

  où $g_1, \dots, g_n$ sont iid $\mathcal{N}(0, 1)$.
- Famille de lois : $\{\mathbb{P}_\theta : \theta \in \mathbb{R}\}$ où

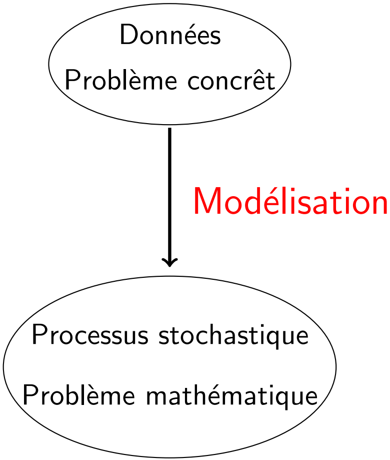
  $$\mathbb{P}_\theta = f_\theta . \lambda^n$$

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

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

  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. Modéle de densité : on observe un n-échantillon

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

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

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

3. modéle 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. :

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

Fonction de répartition empirique et théorème fondamentale de la statistique

Question fondamentale

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

$$X_1, \ldots, X_n$$

qui sont i.i.d. de loi commune $\mathbb{P}_X$.
Rem. : Comme la loi de l'observation $(X_1, \ldots, X_n)$ est $\mathbb{P}_X^{\otimes n}$, se donner un modéle est ici (pour le modéle d'échantillonnage) équivalent à se donner un modéle sur $\mathbb{P}_X$.
Par exemple : $\mathbb{P}_X \in \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$
Une question fondamentale est la suivante :

Question fondamentale

On considère le modéle "total" $=$ $\mathbb{P}_X \in \{$ 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

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

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

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

Formule d'intégration

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

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

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

- Ecriture de $\mathbb{P}^X$:

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

- Formule de calcul (φ fonction test)

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

Loi d'une variable aléatoire (2/4)

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

- La loi de X est décrite par

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

- Ecriture de $\mathbb{P}^X$:

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

- Formule de calcul (φ fonction test)

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

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

- La loi de X est décrite par

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

- Ecriture de $\mathbb{P}^X$:

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

λ : mesure de Lebesgue

- Formule de calcul

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

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 $\mathbb{P}^X$:

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

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

- Formule de calcul

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

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 :

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

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

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

- Désormais, la loi (distribution) 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 $(X_1, \ldots, X_n)$ *:*

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

(C'est une fonction aléatoire)

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

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

$$\boxed{\widehat{F}_n(x) \xrightarrow{p.s.} F(x) \text{ quand } n \to \infty}$$

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

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

Propriétés asymptotiques de $\widehat{F}_n$

Theorem (Glivenko-Cantelli)

$$\left\|\widehat{F}_n - F\right\|_\infty \xrightarrow{p.s.} 0 \; quand \; n \to \infty$$

Aussi appelé Théorème fondamental de la statistique.

Interprétation : Avec un nombre infini de données dans le modéle 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 $\widehat{F}_n(x)$

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

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

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

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

On dit que $\widehat{F}_n(x)$ 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 $n \to \infty$,

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

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 $\sigma(\widehat{F}_n) = \widehat{F}_n(x)^{1/2}(1 - \widehat{F}_n(x))^{1/2}$ (qui est observable), grâce au lemme de Slutsky.

TCL et intervalle de confiance asymptotique

Proposition

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

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

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

$$\mathbb{P}\left[F(x) \in \mathcal{I}_{n,\alpha}^{\mathtt{asymp}}\right] \to 1-\alpha.$$

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 :

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

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

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

- Utile pour le test de Kolmogorov-Smirnov
- version non-asymptotique de ce résultat : quand F est continue, la loi de $\left\|\widehat{F}_n - F\right\|_\infty$ 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 asympotique
2. Un résultat obtenu à n fixé est un résultat dit non-asympotique

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

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

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

On a (Tchebychev)

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

Conduit à

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

Intervalle de confiance

Conclusion : pour tout $\alpha > 0$,

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

Terminologie

L'intervalle

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

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

Inégalité de Hoeffding

Proposition

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

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

<u>Application</u> : 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

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

à comparer avec

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

- Même ordre de grandeur en n.
- Gain significatif dans la limite $\alpha \to 0$. 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)
  $\frac{2}{\sqrt{n}}\widehat{F}_n(x_0)^{1/2}\big(1-\widehat{F}_n(x_0)\big)^{1/2}\Phi^{-1}(1-\alpha/2)$.
- 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 $\alpha \to 0$).

Version non-asymptotique de Kolmogorov-Smirnov

$X_1, \ldots, X_n$ i.i.d. de loi F continue, $\widehat{F}_n$ leur fonction de répartition empirique.

Proposition (Inégalité de Dvoretsky-Kiefer-Wolfowitz)

Pour tout $\varepsilon > 0$.

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

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

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

Rappels de probabilités

Tribus et mesures de probabilité

Soit $\mathfrak{Z}$ un ensemble.

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

   Les éléments de $\mathcal{Z}$ sont appelés des événements.
2. Une mesure de probabilité sur $(\mathfrak{Z}, \mathcal{Z})$ est une appplication $\mathbb{P} : \mathcal{Z} \mapsto [0, 1]$ telle que
   - $\mathbb{P}[\mathfrak{Z}] = 1$
   - Si (A_n) est une famille dénombrable d'événements disjoints alors
   $$\mathbb{P}\left[\cup_n A_n\right] = \sum_n \mathbb{P}[A_n]$$

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

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 pour toute fonction continue bornée $f : \mathbb{R} \mapsto \mathbb{R}$ on a
$$\mathbb{E} f(Z_n) \to \mathbb{E} f(Z)$$
2. (Z_n) converge en probabilité, vers Z, noté $Z_n \xrightarrow{\mathbb{P}} Z$, quand pour tout $\epsilon > 0$,
$$\mathbb{P}\left[|Z_n - Z| \geq \epsilon\right] \to 0$$
3. (Z_n) converge presque surement 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 $\omega \in \Omega_0$
$$Z_n(\omega) \to Z(\omega)$$

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

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

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 surement alors $\mathbb{E}\,|X_1| < \infty$ et elle converge presque surement 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

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

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

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

- 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
$$\mathbb{E}\left[\varphi(X_n, Y_n)\right] \to \mathbb{E}\left[\varphi(X, Y)\right],$$
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 $f(X)$