MAP 433 : Introduction aux méthodes statistiques

31 janvier 2014

Aujourd'hui

1 Agenda

2 Présentation (succinte) du cours

3 Echantillonnage et modélisation statistique (1/2)
- Les données aujourd'hui
- Les données hier...
- Loi d'une variable aléatoire
- Fonction de répartition empirique
- précision d'estimation

Organisation : équipe enseignante

Cours

Marc Hoffmann, Université Paris-Dauphine
hoffmann@ceremade.dauphine.fr

PC

- Stéphane Gaiffas, École Polytechnique.
- Christophe Giraud, Université Paris-Sud et École Polytechnique.
- Guillaume Lecué, École Polytechnique.
- Mathieu Rosenbaum, Université Pierre-et-Marie Curie.

Organisation : agenda

Cours et PC

Les vendredis 31 janvier, 7 février, 14 février, 21 février, 7 mars, 21 mars, 28 mars, 4 avril, 11 avril.

Evaluation

- Contrôle classant : noté sur 20.
- Projet en binôme : (présentés le 7 février) noté sur 4.

Fournit une note tronquée à 20 et convertie en note littérale.

Organisation : materiel

- Transparents du cours téléchargeables à l'adresse

  http ://www.crest.fr/pagesperso.php ?user=3131

- Poly (document autonome contenant l'intégralité du cours et plus, téléchargeable à la même adresse).
- Les documents et exercices de PC.

Présentation (succinte) du cours

- Echantillonnage et modélisation statistique. Expérience statistique (2 cours).
- Méthodes d'estimation classique (2 cours).
- Information statistique, théorie asymptotique pour l'estimation (2 cours).
- Décision statistique et tests (2 cours).
- 1 cours d'ouverture ou de compléments (en fonction de l'auditoire).

Plan

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

Les données aujourd'hui : (1) les chiffres du travail

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/

Les données aujourd'hui : (2)

Le monde de la finance

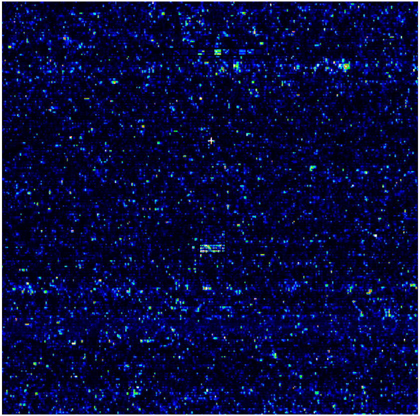
DOW JONES INDUSTRIAL AVERAGE IN (DJI:^DJI)

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/

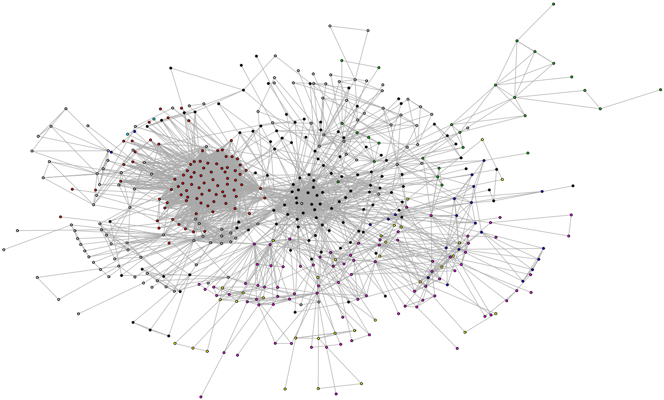
Les données aujourd'hui : (3)

Biopuces et analyse d'ADN



Les données aujourd'hui : (4)

E-marketing - Livres



Retour en arrière : les données hier...

- « Statistik » (dérivé du latin Statisticum). Allemagne, 1740 (Achenwall). Ensemble de mesures et recueil de données nécessaires au fonctionnement et à l'organisation de l'état : recensements et estimations de la population, des richesses, de l'impôt, des armées.
- Les progrès de la statistique : représentation graphique et organisation des données en tableaux (statistique descriptive, dominée par l'école allemande au 18^{e} siècle). Activité importante aussi en Grande Bretagne [1] et dans une France centralisée. [2]

1. William Playfair (1759–1823), 1786, "The Commercial and Political Atlas" contenant le premier diagramme en barres connu.
2. Vauban, 1686, "Méthode générale et facile pour faire le dénombrement des peuples".

Statistique et probabilités

- 17^e siècle : Invention des probabilités. Incorporation d'un raisonnement probabiliste – et donc un modèle du hasard – dans le traitement d'observations.
- Basculement de la statistique vers une discipline scientifique à part entière. Préfigure l'actuariat moderne [3].
- L'exemple historique incontournable : John Arbuthnott (1667–1735) et le déficit des naissances et morts selon le sexe. – la première reflexion « moderne » de statistique.

3. les frêres Hyugens, premier calcul de l'espérance de vie humaine en 1669, Graunt (1620–1674), William Petty (1623–1687), Laplace.

John Arbuthnott et ≪ la divine providence ≫

- 1712, Arbuthnott (médecin de la Reine Anne) examine le nombre de baptêmes de filles et de garçons à Londres, entre 1629 et 1710.
- Sur 82 années retenues, le nombre de naissances masculines est toujours supérieur au nombre de naissance féminines.
- Arbuthnott calcule la probabilité que les naissances masculines (avec équi-probabilité filles/garçons) soient plus nombreuses que les naissances féminines, 82 fois de suite ($= (1/2)^{82}$), ≪ *which will be found easily by the Table of Logarithms to be 1/4 8360 0000 0000 0000 0000 0000* ≫.

An Argument for Divine Providence

- *An Argument for Divine Providence, taken frome the constant Regularity observed in the Births of both Sexes*
- « *[...] This Event is wisely prevented by the Oeconomy of Nature ; and to the judge of the wisdom of the Contrivance, we must observe that the external Accidents to which Males are subject (who must seek their food with danger) do make a great havock of them, and that this loss exceeds far that of the other Sex, occasioned by Disease incident to it, as Experience convinces us. To repair that Loss, provident Nature, by the Disposal of its wife Creator, brings more Males than Females ; and this in almost a constant proportion* » .

Problématique statistique

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

$$\mathtt{x}_1, \ldots, \mathtt{x}_n.$$

- Modélisation statistique :
  - les observations sont des réalisations

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

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

$$\boxed{\left\{\mathbb{P}^n_\vartheta, \vartheta \in \Theta\right\}.}$$

- Problématique : à partir de « l'observation » $X_1, \ldots, X_n$, peut-on retrouver $\mathbb{P}^n_\vartheta$? et donc 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 de $X_1, \ldots, X_n$, établir une décision $\varphi_n(X_1, \ldots, X_n) \in \{\text{ensemble de décisions}\}$ 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 $\bar{X}_{18} = \frac{1}{18}\sum_{i=1}^{18} X_i \overset{\text{ici}}{=} 8/18 = 0.44$. 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

- L'expérience statistique la plus centrale : on observe la réalisation de $X_1, \ldots, X_n$, 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

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

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

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

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

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

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

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

- Formule de calcul

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

Loi d'une variable aléatoire (suite)

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

- La loi de X est décrite par

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

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

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

- Formule de calcul

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

Loi d'une variable aléatoire (suite)

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

- La loi de X est décrite par

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

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

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

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

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

c'est-à-dire

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

Formule de calcul

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

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 :

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

- 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 $\mathbb{P}^X$.

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

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

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

Fonction de répartition empirique

Definition

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

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

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

$$\boxed{\widehat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0), \quad n \to \infty}$$

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

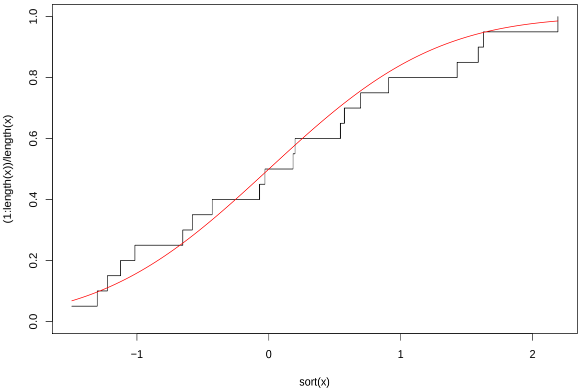


FIGURE : $\widehat{F}_n$ (noir), F (rouge), $n = 20$. $F \sim \mathcal{N}(0, 1)$.

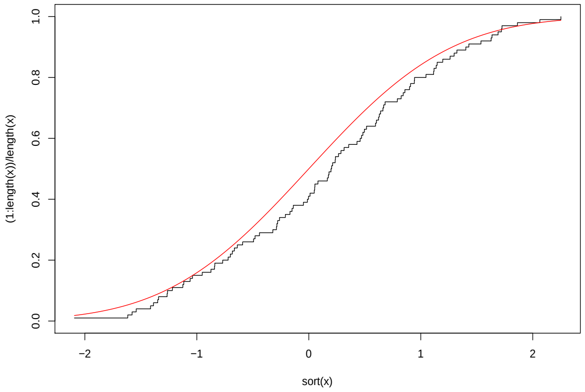


FIGURE : $\widehat{F}_n$ (noir), F (rouge), $n = 100$. $F \sim \mathcal{N}(0,1)$.

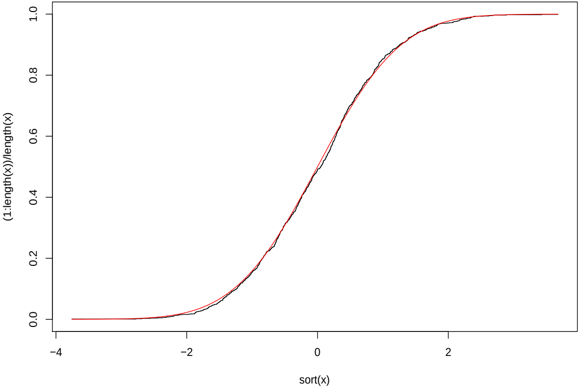


FIGURE : $\widehat{F}_n$ (noir), F (rouge), $n = 1000$. $F \sim \mathcal{N}(0,1)$.

Convergence en probabilité

- Mode de convergence « naturel » en statistique
- Rappel : $X_n \xrightarrow{\mathbb{P}} X$ si

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

- 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 $\forall x_0 \in \mathbb{R}$, $\widehat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$. 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)

Approche naïve : contrôle de la variance

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

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

On a (Tchebychev)

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

Conduit à

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

Intervalle de confiance

<u>Conclusion</u> : pour tout $\alpha > 0$,

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

Terminologie

L'intervalle

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

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

Précision catastrophique !

- Si $\alpha = 5\%$ et $n = 100$, précision $\varepsilon = 0.22$, soit une barre d'erreur de taille 0.44, alors que $0 \leq F(x_0) \leq 1$.
- 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 Hoeffding

Proposition

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

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

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

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

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 ?

L'approche asymptotique

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

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

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

- Outil : Théorème central-limite.

Rappel : théorème central-limite

- TCL :≪ vitesse ≫ dans la loi des grands nombres.
- Si $Y_1, \ldots, Y_n$ i.i.d., $\mu = \mathbb{E}\left[Y_i\right]$, $0 < \sigma^2 = \mathrm{Var}[Y_i] < +\infty$, alors
$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\Big) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$
- Le mode de convergence est la convergence en loi. Ne peut pas avoir lieu en probabilité.
- $X_n \xrightarrow{d} X$ signifie que
$$\mathbb{P}\left[X_n \leq x\right] \to \mathbb{P}\left[X \leq x\right]$$
en tout point x où la fonction de répartition de X est continue (les lois de X_n se ≪ rapprochent ≫ de la loi de X).

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 : $Y_i = 1_{\{X_i \leq x_0\}}$, $\mu = F(x_0)$, $\sigma(F) = F(x_0)^{1/2}\big(1 - F(x_0)\big)^{1/2}$.
  On a

$$\mathbb{P}\left[\left|\widehat{F}_n(x_0) - F(x_0)\right| \geq \varepsilon_n\right] = \mathbb{P}\left[\left|\xi^{(n)}\right| \geq \frac{\sqrt{n}\,\varepsilon_n}{\sigma(F)}\right] = \mathbb{P}\left[\left|\xi^{(n)}\right| \geq \frac{\varepsilon_0}{\sigma(F)}\right]$$

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

TCL et intervalle de confiance (suite)

Il vient

$$\mathbb{P}\left[\left|\xi^{(n)}\right| \geq \frac{\varepsilon_0}{\sigma(F)}\right] \to \int_{|x|\geq \varepsilon_0/\sigma(F)} e^{-x^2/2}\frac{dx}{\sqrt{2\pi}}$$
$$= 2\Big(1 - \Phi\big(\varepsilon_0/\sigma(F)\big)\Big)$$
$$\leq \alpha,$$

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

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

TCL et intervalle de confiance : (suite)

- On a montré

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

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

TCL et intervalle de confiance : conclusion

Proposition

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

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

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

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

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