MAP 433 : Introduction aux méthodes statistiques. Cours 2

4 Septembre 2015

Aujourd'hui

1. Estimation ponctuelle et précision d'estimation
2. Echantillonnage et méthodes empiriques (2/2)
   - Estimation uniforme
   - Estimation de fonctionnelles
3. Modélisation statistique
   - Expérience statistique
   - Expériences dominées
   - Modèle de densité

Cours précédent (rappel)

- A partir de l'observation d'un n-échantillon de loi (de fonction de répartition) inconnue,

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

estimer F.

- Fonction de répartition empirique :

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

- Pour tout $x_0 \in \mathbb{R}$, $\widehat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$ par la loi des grands nombres.

Vers la précision d'estimation

- On a $\forall x_0 \in \mathbb{R}$, $\widehat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$. Questions (auxquelles on a répondu) :
  - 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 (Bienaymé-Techbyshev)
  - non-asymptotique (concentration exponentielle)
  - approche asymptotique (via des théorèmes limites)

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 Hoeffding comparable au TCL en n et α (dans la limite $\alpha \to 0$).

Estimation uniforme

- On « sait » estimer $F(x_0)$, pour un x_0 donné. Qu'en est-il de l'estimation globale de F :

$$\big(F(x), x \in \mathbb{R}\big)?$$

- 3 résultats pour passer de l'estimation en un point à l'estimation globale :
  - Glivenko-Cantelli (convergence uniforme)
  - Kolmogorov-Smirnov (vitesse de convergence, asymptotique)
  - Inégalité de DKW (vitesse de convergence, non-asymptotique)

Glivenko-Cantelli

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

Théorème (Glivenko-Cantelli)

$$\sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right| \stackrel{\text{p.s.}}{\to} 0, \quad \textit{quand } n \to \infty.$$

Glivenko-Cantelli

Supposons que F est continue. Soit $k \in \mathbb{N}$.

- Comme F est continue, il existe des points $-\infty = x_{k,0} < x_{k,1} < \cdots < x_{k,k} = \infty$ tels que $F(x_{k,i}) = i/k$.
- Comme F et $\hat{F}_n$ sont monotones nous avons pour $x \in [x_{k,i-1}, x_{k,i}]$

$$\hat{F}_n(x) - F(x) \leq \hat{F}_n(x_{k,i}) - F(x_{k,i-1}) = \hat{F}_n(x_{k,i}) - F(x_{k,i}) + 1/k$$

et

$$\begin{aligned}\hat{F}_n(x) - F(x) &\geq \hat{F}_n(x_{k,i-1}) - F(x_{k,i}) \\ &= \hat{F}_n(x_{k,i-1}) - F(x_{k,i-1}) - 1/k\,;\end{aligned}$$

ce qui implique

$$\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \sup_i |\hat{F}_n(x_{k,i}) - F(x_{k,i})| + 1/k$$

Glivenko-Cantelli

Supposons que F est continue. Soit $k \in \mathbb{N}$.

- Comme F est continue, il existe des points $-\infty = x_{k,0} < x_{k,1} < \cdots < x_{k,k} = \infty$ tels que $F(x_{k,i}) = i/k$.
- Comme F et $\hat{F}_n$ sont monotones nous avons pour $x \in [x_{k,i-1}, x_{k,i}]$

$$\hat{F}_n(x) - F(x) \leq \hat{F}_n(x_{k,i}) - F(x_{k,i-1}) = \hat{F}_n(x_{k,i}) - F(x_{k,i}) + 1/k$$

et

$$\begin{aligned}\hat{F}_n(x) - F(x) &\geq \hat{F}_n(x_{k,i-1}) - F(x_{k,i}) \\ &= \hat{F}_n(x_{k,i-1}) - F(x_{k,i-1}) - 1/k\,;\end{aligned}$$

ce qui implique

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \leq \sup_i |\hat{F}_n(x_{k,i}) - F(x_{k,i})| + 1/k$$

Kolmogorov-Smirnov

Théorème (Kolmogorov-Smirnov)

Si F est continue,

$$\sqrt{n}\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x)-F(x)\right| \xrightarrow{d} \mathbb{B}, \quad \textit{quand } n\to\infty.$$

$\mathbb{B}$ *v.a. dont la loi est connue et ne dépend pas de F.*

- Difficile.. théorème limite fonctionnel !

Inégalité de DKW

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

Proposition (Inégalité de Dvoretzky-Kiefer-Wolfowitz)

Pour tout $\varepsilon > 0$.

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

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

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

Estimation de fonctionnelles

- Objectif : estimation d'une caractéristique scalaire de la loi inconnue $F \equiv$ estimation d'une fonctionnelle $T(F)$ à valeurs dans $\mathbb{R}$.
- Exemples
  - Déjà vu : valeur en un point $T(F) = F(x_0)$
  - Fonctionnelle régulière :

$$T(F) = h\left(\int_{\mathbb{R}} g(x)dF(x)\right),$$

    où $g, h : \mathbb{R} \to \mathbb{R}$ sont régulières
- Principe méthode de substitution : si $F \rightsquigarrow T(F)$ est "régulière", un estimateur "naturel" est $T(\widehat{F}_n)$

Estimation de fonctionnelles régulières

- Principe : si $F \rightsquigarrow T(F)$ est régulière, alors $T(\widehat{F}_n)$ est un bon estimateur de $T(F)$ (estimateur par substitution).
- Cas où $T(F) = h\big(\int_{\mathbb{R}} g(x)dF(x)\big)$
- Formule de calcul :

$$\boxed{\int_{\mathbb{R}} g(x)d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n g(X_i).}$$

- Estimateur par substitution ou *plug-in* de $T(F)$:

$$\boxed{T(\widehat{F}_n) = h\Big(n^{-1}\sum_{i=1}^n g(X_i)\Big)}$$

Exemples

- Moyenne : $T(F) = m(F) = \int_{\mathbb{R}} x dF(x)$.

$$T(\widehat{F}_n) = m(\widehat{F}_n) = \int_{\mathbb{R}} x d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n.$$

- Variance :

$$T(F) = \sigma^2(F) = \int_{\mathbb{R}} \big(x - m(F)\big)^2 dF(x)$$
$$= \int_{\mathbb{R}} x^2 dF(x) - \Big(\int_{\mathbb{R}} x dF(x)\Big)^2.$$

$$T(\widehat{F}_n) = \sigma^2(\widehat{F}_n) = \int_{\mathbb{R}} \big(x - m(\widehat{F}_n)\big)^2 d\widehat{F}_n(x)$$
$$= \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$$

Exemple

- On modélise les arrivées d'évènements dans un système à évènements discrets (file d'attente, fiabilité)
- On observe $X_1, X_2, \ldots, X_n$ les intervalles de temps entre n arrivées successives.
- Modèle :
  (i) $\{(X_i)\}_{i=1}^n$ sont indépendantes et identiquement distribuées,
  (ii) Les variables X_i sont distribuées suivant une loi exponentielle de paramètre θ, de densité

$$p_\theta(x) = \theta \mathrm{e}^{-\theta x} \mathbb{1}_{\mathbb{R}^+}(x)\,, \quad \theta \in \mathbb{R}^+$$

Absence de mémoire : $\mathbb{P}_\theta(X > x + y | X > x) = \mathbb{P}_\theta(X > y)$

Problème : estimer θ

- Pour tout $k \in \mathbb{N}$, $\mathbb{E}_\theta[X^k] = k!\theta^{-k}$ et donc

$$\theta = \Big(k!/\,\mathbb{E}_\theta[X^k]\Big)^{1/k} = h_k(\mathbb{E}_\theta[X^k])\,.$$

- Estimateur par substitution

$$\hat{\theta}_n = (k!)^{1/k}\left(n^{-1}\sum_{i=1}^n X_i^k\right)^{-1/k}$$

- Meilleur choix de k ?

Exemples de fonctionnelles : quantiles

- Quantiles :

F est continue et strictement croissante $\Longrightarrow$ le **quantile d'ordre** p, $0 < p < 1$, de la loi F est défini comme solution de

$$F(q_p) = p \qquad (\ q_p = F^{-1}(p)\ ).$$

Cas général (F n'est pas strictement $\uparrow$ ou n'est pas continue) :

$$q_p(F) = \tfrac{1}{2}\big(\inf\{x,\ F(x) > p\} + \sup\{x,\ F(x) < p\}\big).$$

La **médiane** :

$$\mathrm{med}(F) = q_{1/2}(F).$$

Les **quartiles** $= \{\mathrm{med}(F), q_{1/4}(F), q_{3/4}(F)\}$.

Quantiles empiriques

Quantile ("théorique") d'ordre p :

$$T(F) = q_p(F) = \frac{1}{2}\big(\inf\{x,\ F(x) > p\} + \sup\{x,\ F(x) < p\}\big).$$

- Avantage : les quantiles sont bien définis **pour toute loi** F.

Quantile empirique d'ordre p :

$$T(\widehat{F}_n) = \widehat{q}_{n,p} = \frac{1}{2}\big(\inf\{x,\ \widehat{F}_n(x) > p\} + \sup\{x,\ \widehat{F}_n(x) < p\}\big).$$

Quantiles empiriques

Expression explicite du quantile empirique d'ordre p :

$$\widehat{q}_{n,p} = \begin{cases} X_{(k)} & \text{si} \quad p \in \big((k-1)/n, k/n\big) \\ \frac{1}{2}\big(X_{(k)} + X_{(k+1)}\big) & \text{si} \quad p = k/n \end{cases}$$

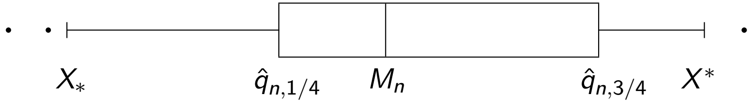
pour $k = 1, \dots, n$, où les $X_{(i)}$ sont **les statistiques d'ordre** associées à l'échantillon $(X_1, \dots, X_n)$:

$$X_{(1)} \leq \cdots \leq X_{(i)} \leq \cdots \leq X_{(n)}.$$

En particulier, la médiane empirique :

$$\boxed{M_n = \mathrm{med}(\widehat{F}_n) = \begin{cases} X_{((n+1)/2)} & \text{pour} \quad n \text{ impair} \\ \frac{1}{2}\big(X_{(n/2)} + X_{(n/2+1)}\big) & \text{pour} \quad n \text{ pair} \end{cases}}$$

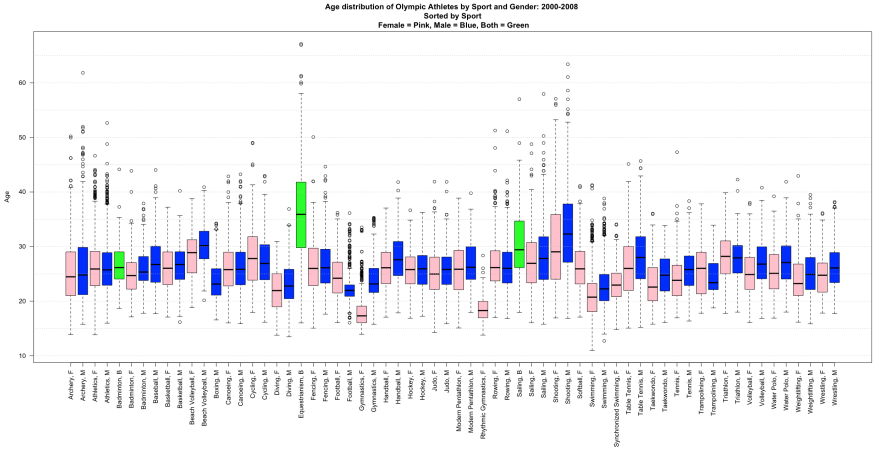
Le boxplot



$$X_* = \min\{X_i : |X_i - \hat{q}_{n,1/4}| \leq 1,5\mathcal{I}_n\},$$
$$X^* = \max\{X_i : |X_i - \hat{q}_{n,3/4}| \leq 1,5\mathcal{I}_n\}.$$

Intervalle interquartile :

$$\mathcal{I}_n = \hat{q}_{n,3/4} - \hat{q}_{n,1/4}.$$


Age distribution of Olympic Athletes by Sport and Gender: 2000-2008
Sorted by Sport
Female = Pink, Male = Blue, Both = Green
Age
10
20
30
40
50
60
Archery, F
Archery, M
Athletics, F
Athletics, M
Badminton, B
Badminton, F
Badminton, M
Baseball, M
Basketball, F
Basketball, M
Beach Volleyball, F
Beach Volleyball, M
Boxing, M
Canoeing, F
Canoeing, M
Cycling, F
Cycling, M
Diving, F
Diving, M
Equestrianism, B
Fencing, F
Fencing, M
Football, F
Football, M
Gymnastics, F
Gymnastics, M
Handball, F
Handball, M
Hockey, F
Hockey, M
Judo, F
Judo, M
Modern Pentathlon, F
Modern Pentathlon, M
Rhythmic Gymnastics, F
Rowing, F
Rowing, M
Sailing, B
Sailing, F
Sailing, M
Shooting, F
Shooting, M
Softball, F
Swimming, F
Swimming, M
Synchronized Swimming, F
Table Tennis, F
Table Tennis, M
Taekwondo, F
Taekwondo, M
Tennis, F
Tennis, M
Trampolining, F
Trampolining, M
Triathlon, F
Triathlon, M
Volleyball, F
Volleyball, M
Water Polo, F
Water Polo, M
Weightlifting, F
Weightlifting, M
Wrestling, F
Wrestling, M

Convergence de l'estimateur par substitution

$$T(F) = h\left(\int_{\mathbb{R}} g(x) dF(x)\right) \quad \text{et} \quad T(\hat{F}_n) = h\left(\int_{\mathbb{R}} g(x) d\hat{F}_n(x)\right)$$

Théorème (Convergence)

si $g, h : \mathbb{R} \to \mathbb{R}$, h continue et $\mathbb{E}\,|g(X)| < \infty$, alors

$$T(\widehat{F}_n) \xrightarrow{\mathbb{P}} T(F)$$

Delta-méthode : cas scalaire

Théorème

Soit $\phi : \mathbb{D}_\phi \subset \mathbb{R} \mapsto \mathbb{R}$ *une fonction définie sur un sous-ensemble ouvert de* $\mathbb{R}$ *et différentiable au point* μ. *Soit* $(T_n)_{n\geq 0}$ *une suite de v.a. à valeurs dans* $\mathbb{D}_\phi$ *et* $(r_n)_{n\geq 0}$ *une suite croissante telle que* $\lim_{n\to\infty} r_n = \infty$. *Si* $r_n(T_n - \mu) \xrightarrow{d} T$, *alors*

$$r_n\{\phi(T_n) - \phi(\mu)\} \xrightarrow{d} \phi'(\mu) T\,.$$

Preuve

- La fonction ϕ est différentiable au point μ ;

$$\phi(t) = \phi(\mu) + \phi'(\mu)(t - \mu) + (t - \mu)\psi(t)\,.$$

Preuve

- La fonction ϕ est différentiable au point μ ;

$$\phi(t) = \phi(\mu) + \phi'(\mu)(t - \mu) + (t - \mu)\psi(t)\,.$$

- Donc,

$$r_n\{\phi(T_n) - \phi(\mu)\} = \phi'(\mu)r_n\{T_n - \mu\} + r_n\{T_n - \mu\}\psi_\mu(T_n)\,.$$

Preuve

- La fonction ϕ est différentiable au point μ ;

$$\phi(t) = \phi(\mu) + \phi'(\mu)(t-\mu) + (t-\mu)\psi(t)\,.$$

- Donc,

$$r_n\{\phi(T_n) - \phi(\mu)\} = \phi'(\mu) r_n\{T_n - \mu\} + r_n\{T_n - \mu\}\psi_\mu(T_n)\,.$$

- Comme ψ est continue en μ, $T_n \xrightarrow{\mathbb{P}} \mu \Rightarrow \psi(T_n) \xrightarrow{\mathbb{P}} \psi(\mu)$.

Preuve

- La fonction ϕ est différentiable au point μ ;

$$\phi(t) = \phi(\mu) + \phi'(\mu)(t-\mu) + (t-\mu)\psi(t)\,.$$

- Donc,

$$r_n\{\phi(T_n) - \phi(\mu)\} = \phi'(\mu) r_n\{T_n - \mu\} + r_n\{T_n - \mu\}\psi_\mu(T_n)\,.$$

- Comme ψ est continue en μ, $T_n \xrightarrow{\mathbb{P}} \mu \Rightarrow \psi(T_n) \xrightarrow{\mathbb{P}} \psi(\mu)$.
- Comme $r_n(T_n - \mu) \xrightarrow{d} T$, on conclut par le Lemme de Slutsky.

Vitesse de convergence de l'estimateur de substitution : Etape 1

$$T(F) = h\Big(\int_{\mathbb{R}} g(x)dF(x)\Big) \quad \text{et} \quad T(\widehat{F}_n) = h\Big(\int_{\mathbb{R}} g(x)d\widehat{F}_n(x)\Big)$$

- TCL :

$$\sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \int_{\mathbb{R}} g(x)dF(x)\Big) \xrightarrow{d} \mathcal{N}\big(0, \mathrm{Var}\big[g(X)\big]\big),$$

où X est une v.a. de loi F et

$$\begin{aligned}\mathrm{Var}\big[g(X)\big] &= \mathbb{E}\left[g(X)^2\right] - \big(\mathbb{E}[g(X)]\big)^2 \\ &= \int_{\mathbb{R}} g(x)^2 dF(x) - \Big(\int_{\mathbb{R}} g(x)dF(x)\Big)^2.\end{aligned}$$

Vitesse de convergence de l'estimateur de substitution : Etape 2

$$T(F) = h\left(\int_{\mathbb{R}} g(x)dF(x)\right) \quad \text{et} \quad T(\hat{F}_n) = h\left(\int_{\mathbb{R}} g(x)d\hat{F}_n(x)\right)$$

- TCL :

$$\sqrt{n}\Big(n^{-1}\sum_{i=1}^{n} g(X_i) - \int_{\mathbb{R}} g(x)dF(x)\Big) \xrightarrow{d} \mathcal{N}\left(0, \mathrm{Var}\big[g(X)\big]\right),$$

- Delta-méthode

$$\sqrt{n}\{T(F_n) - T(F)\} \xrightarrow{d} \mathcal{N}\left(0, \{h'(\mathbb{E}[g(X)])\}^2 \,\mathrm{Var}[g(X)]\right)$$

car si $Z \sim \mathcal{N}(0,\sigma^2)$, $aZ \sim \mathcal{N}(0, a^2\sigma^2)$.

Conclusion

Proposition

Si $\mathbb{E}[g(X)^2] < +\infty$ *et h continûment différentiable, alors*

$$\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big) \xrightarrow{d} \mathcal{N}\big(0, v(F)\big),$$

où $v(F) = h'\big(\mathbb{E}\left[g(X)\right]\big)^2 \mathrm{Var}\big[g(X)\big]$.

Pour construire un intervalle de confiance, il faut encore remplacer $v(F)$ par $v(\widehat{F}_n)$. On montre que $v(\widehat{F}_n) \xrightarrow{\mathbb{P}} v(F)$ et, via le lemme de Slutsky,

$$\sqrt{n}\frac{T(\widehat{F}_n) - T(F)}{v(\widehat{F}_n)^{1/2}} \xrightarrow{d} \mathcal{N}(0,1).$$

On en déduit un intervalle de confiance asymptotique comme précédemment.

Le cas de la dimension $d > 1$

- Il s'agit de fonctionnelles de la forme

$$T(F) = h\left(\int_{\mathbb{R}} g_1(x)dF(x), \ldots, \int_{\mathbb{R}} g_k(x)dF(x)\right)$$

où $h : \mathbb{R}^k \to \mathbb{R}$ continûment différentiable.

- Exemple : le coefficient d'asymétrie

$$T(F) = \frac{\int_{\mathbb{R}} \big(x - m(F)\big)^3 dF(x)}{\sigma^{3/2}(F)},$$

$m(F) =$ moyenne de F, $\sigma^2(F) =$ variance de F.

- Outil : Version multidimensionnelle du TCL et de la « méthode delta ».

Méthode « delta » multidimensionnelle

- TCL multidimensionnel : $(\boldsymbol{X}_n)_{n\geq 1}$ vecteurs aléatoires dans $\mathbb{R}^k$, i.i.d., de moyenne $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_1]$ et de matrice de variance-covariance $\Sigma = \mathbb{E}\left[(\boldsymbol{X}_1 - \boldsymbol{\mu})(\boldsymbol{X}_1 - \boldsymbol{\mu})^T\right]$ bien définie. Alors $\bar{\boldsymbol{X}}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i$ vérifie :

$$\sqrt{n}\big(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\big) \xrightarrow{d} \mathcal{N}\big(0, \Sigma\big).$$

- Méthode « delta » multidimensionnelle : Si, de plus, $h : \mathbb{R}^k \to \mathbb{R}$ continûment différentiable, alors

$$\sqrt{n}\big(h(\overline{\boldsymbol{X}}_n) - h(\boldsymbol{\mu})\big) \xrightarrow{d} \mathcal{N}\Big(0, \nabla h(\boldsymbol{\mu})\Sigma\nabla h(\boldsymbol{\mu})^T\Big).$$

Application : coefficient d'asymétrie

- Coefficient d'asymétrie : on a

$$T(F) = h\Big(\int_{\mathbb{R}} x dF(x), \int_{\mathbb{R}} x^2 dF(x), \int_{\mathbb{R}} x^3 dF(x)\Big)$$

avec

$$h(\alpha, \beta, \gamma) = \frac{\gamma - 3\alpha\beta + 2\alpha^3}{(\beta - \alpha^2)^{3/2}}.$$

$$T(\widehat{F}_n) = h\Big(\tfrac{1}{n}\sum_{i=1}^n X_i, \tfrac{1}{n}\sum_{i=1}^n X_i^2, \tfrac{1}{n}\sum_{i=1}^n X_i^3\Big).$$

- On applique le TCL multidimensionnel avec $\boldsymbol{X}_i = (X_i, X_i^2, X_i^3)^T$ et $\boldsymbol{\mu} = \big(\int_{\mathbb{R}} x dF(x), \int_{\mathbb{R}} x^2 dF(x), \int_{\mathbb{R}} x^3 dF(x)\big)^T$, puis la méthode « delta » avec h.

Limites de l'approche empirique

L'estimation de $T(F)$ par $T(\widehat{F}_n)$ n'est pas toujours possible :

- La fonctionnelle $F \rightsquigarrow T(F)$ n'est pas « régulière »,
- La paramétrisation $F \rightsquigarrow T(F)$ ne donne pas lieu à une forme analytique simple. $\rightarrow$ autres approches.

Exemple. Hypothèse : F admet une densité f par rapport à le mesure de Lebesgue, continue ($=$ pp à une fonction continue f).

$$T(F) = f(x_0), \;\; x_0 \in \mathbb{R} \;\; \text{(donné)}.$$

On ne peut pas prendre comme estimateur $\widehat{F}_n'(x_0)$ car $\widehat{F}_n$ n'est pas différentiable (constante par morceaux...)

Limites de l'approche empirique

L'estimation de $T(F)$ par $T(\widehat{F}_n)$ n'est pas toujours souhaitable :

- Souvent on dispose d'information a priori supplémentaire : F appartient à une sous-classe particulière de distributions, et il y a des choix plus judicieux que l'estimateur par substitution.

Conclusion

- L'approche empirique, basée sur $\widehat{F}_n$ permet d'estimer une distribution inconnue F ou une fonctionnelle $T(F) \in \mathbb{R}$ à partir d'un n-échantillon, mais
  - reste très générale, pas toujours adaptée.
  - restreinte à la situation d'un n-échantillon.
- Formalisation de la notion d'expérience statistique
  - incorporation d'information de modélisation supplémentaire.
  - construction de méthodes d'estimation – de décision – systématiques.
  - comparaison et optimalité des méthodes.

Expérience statistique

Consiste à identifier :

- Des observations
$$\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_n$$
considérées comme des réalisations de variables aléatoires $Z = (X_1, \ldots, X_n)$ de loi $\mathbb{P}^Z$.
- Une famille de lois
$$\{\mathbb{P}_\theta,\ \theta \in \Theta\}.$$
- Une problématique "estimer" le paramètre θ ou bien prendre une décision sur une propriété relative à θ (test).

Expérience statistique

- Approche générale empirique :
  - $\theta = F$, Θ est l'ensemble de toutes les lois (s'il s'agit de l'estimation de F) ;
  - $\theta = F$, Θ est l'ensemble de toutes les lois vérifiant une hypothèse très générale, par exemple, la bornitude d'un moment (s'il s'agit de l'estimation de $T(F)$).
- Approche paramétrique : on suppose que F appartient à une famille de lois connue indexée par un paramètre θ de dimension finie : $\theta \in \Theta \subset \mathbb{R}^d$.
  - Exemple : $\Theta = \mathbb{R}$,

$$X_i = \theta + \xi_i, \quad i = 1, \ldots, n,$$

ξ_i v.a. i.i.d. de densité connue f sur $\mathbb{R}$ et $\mathbb{E}(X_i) = \theta$.
Question : en utilisant cette information supplémentaire, peut-on construire un estimateur plus performant que l'estimateur $\bar{X}_n$ basé sur l'approche empirique ?

Expérience statistique

- En écrivant
$$X_i = \theta + \xi_i, \quad i = 1, \ldots, n,$$
ξ_i v.a. i.i.d. de densité connue f, nous précisons la forme de la loi $\mathbb{P}_\theta$ de $(X_1, \ldots, X_n)$:
$$\mathbb{P}_\theta\left[A\right] = \int_A \left(\prod_{i=1}^n f(x_i - \theta)\right) dx_1 \ldots dx_n,$$
pour tout $A \in \mathcal{B}(\mathbb{R}^n)$.

Expérience statistique

Definition

Une expérience (un modèle) statistique $\mathcal{E}$ est le triplet

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

avec

- $(\mathfrak{Z}, \mathcal{Z})$ *espace mesurable (souvent $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$),*
- $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ *famille de probabilités définies simultanément sur le même espace $(\mathfrak{Z}, \mathcal{Z})$,*
- θ *est le paramètre (inconnu), et Θ est l'ensemble des paramètres.*

Experience engendrée par $(X_1, \ldots, X_n)$

- On observe

$$Z = (X_1, \ldots, X_n), \qquad X_i = \theta + \xi_i,$$

ξ_i v.a. i.i.d. de densité connue f.

- La famille de lois $\{\mathbb{P}_\theta^n, \theta \in \Theta = \mathbb{R}\}$ est définie sur $\mathfrak{Z} = \mathbb{R}^n$ par

$$\mathbb{P}_\theta^n[A] = \int_A \left(\prod_{i=1}^n f(x_i - \theta)\right) dx_1 \ldots dx_n,$$

pour $A \in \mathcal{Z} = \mathcal{B}(\mathbb{R}^n)$ (et $\mathbb{P}^Z$ est l'une des $\mathbb{P}_\theta^n$).

- Expérience engendrée par l'observation Z :

$$\mathcal{E}^n = \left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \{\mathbb{P}_\theta^n, \theta \in \Theta\}\right).$$

Expérience (modèle) paramétrique, non-paramétrique

- Si Θ est un sous-ensemble de $\mathbb{R}^d$: expérience (=modèle) paramétrique.
- Sinon (par exemple si le paramètre θ est un élément d'un espace fonctionnel) : expérience (=modèle) non-paramétrique.

Expériences dominées

- On fait une hypothèse minimale de « complexité » sur le modèle statistique. But : ramener l'étude de la famille

$$\{\mathbb{P}_\theta,\, \theta \in \Theta\}$$

à l'étude d'une famille de fonctions

$$\{z \in \mathfrak{Z} \rightsquigarrow f(\theta, z) \in \mathbb{R}_+,\, \theta \in \Theta\}.$$

- Via la notion de domination. Si μ, ν sont deux mesures σ-finies sur $\mathfrak{Z}$, alors μ domine ν (notation $\nu \ll \mu$) si

$$\mu[A] = 0 \Rightarrow \nu[A] = 0.$$

Théorème de Radon-Nikodym

Théorème

Si $\nu \ll \mu$, il existe une fonction positive

$$z \rightsquigarrow p(z) \stackrel{notation}{=} \frac{d\nu}{d\mu}(z),$$

définie $\mu-$p.p., $\mu-$ intégrable, telle que

$$\nu\big[A\big] = \int_A p(z)\mu(dz) = \int_A \tfrac{d\nu}{d\mu}(z)\mu(dz), \quad A \in \mathcal{Z}.$$

Expérience dominée

Definition

Une expérience statistique $\mathcal{E} = \big(\mathfrak{Z}, \mathcal{Z}, \{\, \mathbb{P}_\theta, \theta \in \Theta \}\big)$ est dominée par la mesure σ-finie μ définie sur $\mathfrak{Z}$ si

$$\forall \theta \in \Theta : \mathbb{P}_\theta \ll \mu.$$

On appelle densités de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ la famille de fonctions (définies $\mu-$ p.p.)

$$z \rightsquigarrow \frac{d\,\mathbb{P}_\theta}{d\mu}(z),\ z \in \mathfrak{Z},\ \theta \in \Theta.$$

Modèle de densité (paramétrique)

- On observe un n-échantillon de v.a.r. $X_1, \ldots, X_n$.
- La loi des X_i appartient à $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$, famille de probabilités sur $\mathbb{R}$, dominée par une mesure (σ-finie) $\mu(dx)$ sur $\mathbb{R}$.
- La loi de $(X_1, \ldots, X_n)$ s'écrit

$$\begin{aligned}\mathbb{P}_\theta^n(dx_1 \cdots dx_n) &= \mathbb{P}_\theta(dx_1) \otimes \cdots \otimes \mathbb{P}_\theta(dx_n) \\ &\ll \mu(dx_1) \otimes \cdots \otimes \mu(dx_n) \\ &\overset{\text{notation}}{=} \mu^n(dx_1 \cdots dx_n)\end{aligned}$$

Modèle de densité (paramétrique)

- Densité du modèle : on part de

$$f(\theta, x) = \frac{d\,\mathbb{P}_\theta}{d\mu}(x), \quad x \in \mathbb{R}$$

et

$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^n}(x_1, \ldots, x_n) = \prod_{i=1}^n f(\theta, x_i), \quad x_1, \ldots, X_n \in \mathbb{R}\,.$$

- L'expérience statistique engendrée par $(X_1, \ldots, X_n)$ s'écrit :

$$\mathcal{E}^n = \Big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \big\{\mathbb{P}_\theta^n, \theta \in \Theta\big\}\Big), \quad \Theta \subset \mathbb{R}^d\,.$$

Exemple 1 : modèle de densité gaussienne univariée

- $X_i \sim \mathcal{N}(m, \sigma^2)$, avec
$$\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}.$$
$$\mathbb{P}_\theta(dx) = f(\theta, x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-m)^2}{2\sigma^2}\Big) dx$$
$$\ll \mu(dx) = dx.$$
- Puis
$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^n}(x_1, \ldots, x_n) = \prod_{i=1}^n f(\theta, x_i)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - m)^2\Big),$$

avec $x_1, \ldots, x_n \in \mathbb{R}$.

Exemple 2 : modèle de Bernoulli

- $X_i \sim \text{Bernoulli}(\theta)$, avec $\theta \in \Theta = [0, 1]$.

$$\mathbb{P}_\theta(dx) = (1 - \theta)\,\delta_0(dx) + \theta\,\delta_1(dx)$$
$$\ll \mu(dx) = \delta_0(dx) + \delta_1(dx) \;\; \text{(mesure de comptage)}.$$

- Puis

$$\boxed{\frac{d\,\mathbb{P}_\theta}{d\mu}(x) = (1 - \theta)\,1_{\{x=0\}} + \theta\,1_{\{x=1\}} = \theta^x (1 - \theta)^{1-x}}$$

avec $x \in \{0, 1\}$, et

$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^n}(x_1, \cdots, x_n) = \prod_{i=1}^n \theta^{x_i}(1 - \theta)^{1-x_i},$$

avec $x_i \in \{0, 1\}$.

Exemple 3 : temps de panne arrêtés

- On observe $X_1, \ldots, X_n$, où $X_i = Y_i \wedge T$, avec Y_i lois exponentielles de paramètre θ et T temps fixe (censure).
- Cas 1 : $T = \infty$ (pas de censure). Alors $\theta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et

$$\mathbb{P}_\theta(dx) = \theta \exp(-\theta x) 1_{\{x \geq 0\}} dx \ll \mu(dx) = dx$$

et

$$\frac{d\,\mathbb{P}_\theta^n}{d\mu^n}(x_1, \ldots, x_n) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right),$$

avec $x_i \in \mathbb{R}_+$.

- Cas 2 : Comment s'écrit le modèle dans la cas où $T < \infty$ (présence de censure) ? Comment choisir μ ?