

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,

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

estimer F .

- Fonction de répartition empirique :

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

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

Vers la précision d'estimation

- On a $\forall x_0 \in \mathbb{R}, \hat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$. 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-Tchbyshev)
 - 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)
 $\frac{2}{\sqrt{n}} \hat{F}_n(x_0)^{1/2} (1 - \hat{F}_n(x_0))^{1/2} \Phi^{-1}(1 - \alpha/2).$
- La longueur la plus petite est (**sans surprise !**) celle fournie par le TCL. Mais Hoeffding **comparable** au TCL en n et α (dans la limite $\alpha \rightarrow 0$).

Estimation uniforme

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

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

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

Théorème (Glivenko-Cantelli)

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0, \quad \text{quand } n \rightarrow \infty.$$

Glivenko-Cantelli

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

- Comme F est continue, il existe des points $-\infty = x_{k,0} < x_{k,1} < \dots < x_{k,k} = \infty$ 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 $-\infty = x_{k,0} < x_{k,1} < \dots < x_{k,k} = \infty$ 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,

$$\sqrt{n} \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \xrightarrow{d} \mathbb{B}, \quad \text{quand } n \rightarrow \infty.$$

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

- Difficile.. théorème limite fonctionnel !

Inégalité de DKW

X_1, \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 :

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

Estimation de fonctionnelles

- **Objectif** : estimation d'une caractéristique scalaire de la loi inconnue $F \equiv$ estimation d'une fonctionnelle $F \rightsquigarrow T(F)$ à valeurs réelles.
- **Exemples**
 - valeur en un point $T(F) = F(x_0)$
 - $T(F) = h\left(\int_{\mathbb{R}} g(x)dF(x)\right)$, où $g, h : \mathbb{R} \rightarrow \mathbb{R}$ sont régulières
 - Autres cas...
- Principe **méthode de substitution** : si $F \rightsquigarrow T(F)$ est "régulière", un estimateur "naturel" est $T(\hat{F}_n)$

Exemple

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

- Formule de calcul :

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

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

$$T(\hat{F}_n) = h\left(n^{-1} \sum_{i=1}^n g(X_i)\right)$$

Exemples

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

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

- Variance :

$$\begin{aligned} T(F) = \sigma^2(F) &= \int_{\mathbb{R}} (x - m(F))^2 dF(x) \\ &= \int_{\mathbb{R}} x^2 dF(x) - \left(\int_{\mathbb{R}} x dF(x) \right)^2. \end{aligned}$$

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

Exemple

- On modélise les arrivées d'évènements dans un système à évènements discrets (file d'attente, fiabilité)
- On observe X_1, X_2, \dots, 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

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

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

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

- 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** \implies le **quantile d'ordre** p , $0 < p < 1$, de la loi F est défini comme solution de

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

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

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

La **médiane** :

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

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

Quantiles empiriques

Quantile ("théorique") d'ordre p :

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

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

Quantile empirique d'ordre p :

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

Quantiles empiriques

Expression explicite du quantile empirique d'ordre p :

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

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

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

En particulier, la médiane empirique :

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

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 :

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



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

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

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}_ϕ et $(r_n)_{n \geq 0}$ une suite croissante telle que $\lim_{n \rightarrow \infty} r_n = \infty$. 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 ;

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

Preuve

- La fonction ϕ est différentiable au point μ ;

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

- 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 ;

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

- 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 ;

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

- 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

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

■ TCL :

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

où

$$\text{Var}[g(X)] = \int_{\mathbb{R}} g(x)^2 dF(x) - \left(\int_{\mathbb{R}} g(x) dF(x) \right)^2.$$

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 :

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

■ Delta-méthode

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

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

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

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

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

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

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

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

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

- **Exemple** : le coefficient d'asymétrie

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

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

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

Méthode « delta » multidimensionnelle

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

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

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

$$\sqrt{n}(h(\bar{\mathbf{X}}_n) - h(\boldsymbol{\mu})) \xrightarrow{d} \mathcal{N}\left(0, \nabla h(\boldsymbol{\mu}) \Sigma \nabla h(\boldsymbol{\mu})^T\right).$$

Application : coefficient d'asymétrie

- Coefficient d'asymétrie : on a

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

avec

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

$$T(\hat{F}_n) = h\left(\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n X_i^2, \frac{1}{n} \sum_{i=1}^n X_i^3\right).$$

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

Limites de l'approche empirique

L'estimation de $T(F)$ par $T(\hat{F}_n)$ 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**. → autres approches.

Exemple. **Hypothèse** : F admet une densité f par rapport à la mesure de Lebesgue, **continue** (= presque-partout à une fonction continue f).

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

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

Limites de l'approche empirique

L'estimation de $T(F)$ par $T(\hat{F}_n)$ 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 \hat{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

$$x_1, x_2, \dots, x_n$$

considérées comme des réalisations de variables aléatoires $Z = (X_1, \dots, X_n)$ 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 : $\Theta = \mathbb{R}$,

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

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

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

ξ_i v.a. i.i.d. de densité connue f , nous précisons la forme de la loi \mathbb{P}_θ de (X_1, \dots, X_n) :

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

pour tout $A \in \mathcal{B}(\mathbb{R}^n)$.

Expérience statistique

Definition

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

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

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*.

Expérience engendrée par (X_1, \dots, X_n)

- On observe

$$Z = (X_1, \dots, X_n), \quad X_i = \theta + \xi_i,$$

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 $\mathcal{Z} = \mathbb{R}^n$ par

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

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

- Expérience **engendrée par l'observation** Z :

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

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

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

définie μ -p.p., μ -intégrable, telle que

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

Expérience dominée

Definition

Une expérience statistique $\mathcal{E} = (\mathfrak{Z}, \mathcal{Z}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ 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 μ -p.p.)

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

Modèle de densité (paramétrique)

- On observe un n -échantillon de v.a.r. X_1, \dots, 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, \dots, X_n) s'écrit

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

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

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

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

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

Exemple 1 : modèle de densité gaussienne univariée

- $X_i \sim \mathcal{N}(m, \sigma^2)$, avec

$$\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+ \setminus \{0\}.$$

$$\mathbb{P}_\theta(dx) = f(\theta, x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)dx \\ \ll \mu(dx) = dx.$$

- Puis

$$\frac{d\mathbb{P}_\theta^n}{d\mu^n}(x_1, \dots, x_n) = \prod_{i=1}^n f(\theta, x_i) \\ = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m)^2\right),$$

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

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

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

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

Exemple 3 : temps de panne arrêtés

- On observe X_1, \dots, 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

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

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 ?