

MAP 433 : Introduction aux méthodes statistiques. Cours 2

7 février 2014

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



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 :

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



Convergence en probabilité

- Mode de convergence « naturel » en statistique
- **Rappel** : $X_n \xrightarrow{\mathbb{P}} X$ si
$$\forall \varepsilon > 0, \mathbb{P}[|X_n - X| \geq \varepsilon] \rightarrow 0, \quad n \rightarrow \infty.$$
- **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...).

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



Vers la précision d'estimation

- On a $\forall x_0 \in \mathbb{R}, \hat{F}_n(x_0) \xrightarrow{\mathbb{P}} F(x_0)$. 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)**

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

◀ ▶ ⏪ ⏩ 🔍 ↺

Approche naïve : contrôle de la variance

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

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

On a (**Tchebychev**)

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

Conduit à

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

◀ ▶ ⏪ ⏩ 🔍 ↺

Intervalle de confiance

Conclusion : pour tout $\alpha > 0$,

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

Terminologie

L'intervalle

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

◀ ▶ ⏪ ⏩ 🔍 ↺

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 ?

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

◀ ▶ ⏪ ⏩ 🔍 ↺

Inégalité de Hoeffding

Proposition

Y_1, \dots, 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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Comparaison Tchebychev vs. Hoeffding

Nouvel intervalle de confiance

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

à comparer avec

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

- Même ordre de grandeur en n .
- Gain **significatif** dans la limite $\alpha \rightarrow 0$. La « prise de risque » devient marginale par rapport au nombre d'observations.
- **Optimalité d'une telle approche ?**

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

L'approche asymptotique

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

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

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

- Outil : **Théorème central-limite**.

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Rappel : théorème central-limite

- TCL : « vitesse » dans la loi des grands nombres.
- Si Y_1, \dots, Y_n i.i.d., $\mu = \mathbb{E}[Y_i]$, $0 < \sigma^2 = \text{Var}[Y_i] < +\infty$, alors

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Y_i - \mu \right) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

- Le mode de convergence est **la convergence en loi**. Ne peut pas avoir lieu en probabilité.
- $X_n \xrightarrow{d} X$ signifie que

$$\mathbb{P}[X_n \leq x] \rightarrow \mathbb{P}[X \leq x]$$

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).

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Interprétation et application

- Interprétation du TCL :

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

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

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



TCL et intervalle de confiance (suite)

Il vient

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

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



TCL et intervalle de confiance : (suite)

- On a montré

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



TCL et intervalle de confiance : conclusion

Proposition

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

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

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

$$\mathbb{P} [F(x_0) \in \mathcal{I}_{n,\alpha}^{\text{asympt}}] \rightarrow 1 - \alpha.$$

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



Lemme de Slutsky

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

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

pour φ continue bornée.

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



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)**
 $\frac{2}{\sqrt{n}} \widehat{F}_n(x_0)^{1/2} (1 - \widehat{F}_n(x_0))^{1/2} \Phi^{-1}(1 - \alpha/2)$.
- 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 $\alpha \rightarrow 0$).

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique



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)

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



Glivenko-Cantelli, Kolmogorov-Smirnov

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

Proposition

- (Glivenko-Cantelli)

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

- (Kolmogorov-Smirnov) Si F est continue,

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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 :

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

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 $g, h : \mathbb{R} \rightarrow \mathbb{R}$ sont **régulières**

- **Principe (méthode de substitution)** : si $F \rightsquigarrow T(F)$ est « régulière », un estimateur « naturel » est $T(\hat{F}_n)$ (**estimateur par plug-in**).

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

Estimation de fonctionnelles régulières

- Cas où $T(F) = h(\int_{\mathbb{R}} g(x) dF(x))$
- Formule de calcul :

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

Traduction : **une variable aléatoire de loi \hat{F}_n prend les valeurs X_i avec probabilité $1/n$.**

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

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 :

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

Exemples

- Asymétrie (skewness) :

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

- Aplatissement (kurtosis) :

$$T(F) = \kappa(F) = \frac{\int_{\mathbb{R}} (x - m(F))^4 dF(x)}{\sigma^2(F)^2} = \dots$$

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



Quantiles empiriques

Quantile ("théorique") d'ordre p :

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

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

Quantile empirique d'ordre p :

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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 :

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

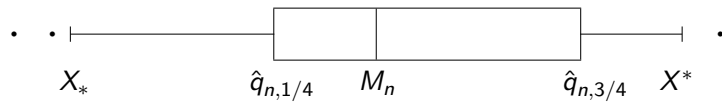
Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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 :

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

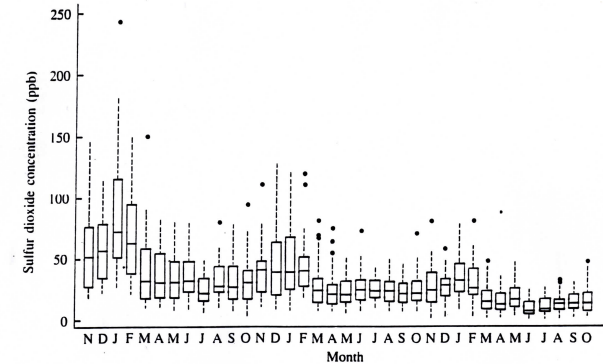
Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

Exemple d'application du boxplot



Boxplots of daily maximum concentrations of sulfur dioxide.

Performance de l'estimateur par substitution

■ **Convergence** si $g, h : \mathbb{R} \rightarrow \mathbb{R}$, h continue et $\mathbb{E} |g(X)| < \infty$, alors $T(\hat{F}_n) \xrightarrow{p.s.} T(F)$ (loi forte des grands nombres).

■ **Vitesse de convergence, Etape 1.** TCL :

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

où X est une v.a. de loi F et

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

Vitesse de convergence (suite)

■ **Etape 2.** On a $\sqrt{n}(Z_n - c_1) \xrightarrow{d} \mathcal{N}(0, c_2)$. Comment transférer ce résultat à $\sqrt{n}(h(Z_n) - h(c_1)) \xrightarrow{d} ?$

■ **Méthode « delta »** : si h continûment différentiable

$$\sqrt{n}(h(Z_n) - h(c_1)) = \sqrt{n}(Z_n - c_1)h'(\eta_n), \quad \eta_n \in [Z_n, c_1].$$

On a $\sqrt{n}(Z_n - c_1) \xrightarrow{d} \mathcal{N}(0, c_2)$ et $h'(\eta_n) \xrightarrow{\mathbb{P}} h'(c_1)$.

Lemme de Slutsky :

$$\sqrt{n}(Z_n - c_1)h'(\eta_n) \xrightarrow{d} \mathcal{N}(0, c_2)h'(c_1).$$

Finalement

$$\boxed{\sqrt{n}(h(Z_n) - h(c_1)) \xrightarrow{d} \mathcal{N}(0, c_2[h'(c_1)]^2)}$$

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique

Conclusion

Proposition

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

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

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 $v(\widehat{F}_n)$. 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.

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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 ».

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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

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

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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 (= pp 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...)

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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 plug-in.

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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.

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Estimation
uniforme
Estimation de
fonctionnelles

Modélisation
statistique



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

$$\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}.$$

- Une problématique : retrouver le paramètre ϑ tel que $\mathbb{P}^Z = \mathbb{P}_\vartheta$ (estimation) ou bien prendre une décision sur une propriété relative à ϑ (test).

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité



Expérience statistique

- Approche générale empirique :
 - $\vartheta = F$, Θ est l'ensemble de toutes les lois (s'il s'agit de l'estimation de F);
 - $\vartheta = F$, 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 : $\vartheta \in \Theta \subset \mathbb{R}^d$.
 - **Exemple** : $\Theta = \mathbb{R}$,

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité



Expérience statistique

- En écrivant

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

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) :

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité



Expérience statistique

Définition

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

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

avec

- $(\mathfrak{Z}, \mathcal{Z})$ espace mesurable (souvent $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$),
- $\{ \mathbb{P}_\vartheta, \vartheta \in \Theta \}$ 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 connus**.

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité



Expérience engendrée par (X_1, \dots, X_n)

- **Traitement sur un exemple** : on observe

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

ξ_i v.a. i.i.d. de densité **connue** f .

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

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

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** :

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité



Expérience (modèle) paramétrique, non-paramétrique

- Si Θ peut être « pris » comme 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.**

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité

Expériences dominées

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

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

à l'étude d'une famille de fonctions

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

- 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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité

Expérience dominée

Définition

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

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

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité

Densité, régression

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Deux classes d'expériences statistiques **dominées** fondamentales :

- Le modèle de **densité** (Cours 3)
- Le modèle de **régression** (Cours 4)

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
**Expériences
dominées**
Modèle de
densité



Modèle de densité (paramétrique)

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

- On observe un n -échantillon de v.a.r. X_1, \dots, X_n .
- La loi des X_i appartient à $\{\mathbb{P}_\vartheta, \vartheta \in \Theta\}$, 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

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

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
**Expériences
dominées**
Modèle de
densité



Modèle de densité (paramétrique)

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

- **Densité du modèle** : on part de

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

et

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

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

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

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
**Expériences
dominées**
Modèle de
densité



Exemple 1 : modèle de densité gaussienne univariée

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

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

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

- Puis

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

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

MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
**Expériences
dominées**
Modèle de
densité



Exemple 2 : modèle de Bernoulli

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

$$\mathbb{P}_\vartheta(dx) = (1 - \vartheta) \delta_0(dx) + \vartheta \delta_1(dx)$$

$$\ll \mu(dx) = \delta_0(dx) + \delta_1(dx) \quad (\text{mesure de comptage}).$$

- Puis

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

avec $x \in \{0, 1\}$ (et 0 sinon), et

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

avec $x_i \in \{0, 1\}$ (et 0 sinon).



MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité

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 $\vartheta \in \Theta = \mathbb{R}_+ \setminus \{0\}$ et

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

et

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

avec $x_i \in \mathbb{R}_+$ (et 0 sinon).

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



MAP 433 :
Introduction
aux méthodes
statistiques.
Cours 2

Estimation
ponctuelle et
précision
d'estimation

Echantillonnage
et méthodes
empiriques
(2/2)

Modélisation
statistique

Expérience
statistique
Expériences
dominées
Modèle de
densité