Processus Empiriques

Master 2 Probabilités et Statistiques

RAHMANI Naceur

24/11/2025

Période Fondatrice (1933-1952)

- **1933**: Glivenko-Cantelli - Convergence uniforme de F_n vers F
- **1933**: Kolmogorov - Distribution limite de $\sup |F_n - F|$
- **1952**: Donsker - Invariance fonctionnelle (TCL fonctionnel)

Développements Théoriques (1960-1980)

- **Années 60**: Généralisation aux processus indexés
- **1971**: Vapnik-Chervonenkis - Théorie VC et entropie
- **1978**: Dudley - TCL uniforme et conditions d'entropie

Période Moderne (1990-Présent)

- **1996**: van der Vaart & Wellner - Synthèse complète
- **Années 2000**: Applications en apprentissage automatique
- **Aujourd'hui**: Extensions aux données haute dimension

Cadre Probabiliste

- Espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$
- Données: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} P$
- Mesure empirique:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$$

Définition Générale

$$G_n(f) = \sqrt{n}(P_n f - P f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X)])$$

pour $f \in \mathcal{F}$ (classe de fonctions mesurables).

Interprétation

- Fluctuation normalisée de la mesure empirique
- Généralisation multidimensionnelle du TCL

Cadre Spécifique

- Classe de fonctions: $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}, t \in \mathbb{R}\}$
- Fonction de répartition: $F(t) = P(X \leq t)$
- Fonction de répartition empirique:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}$$

Définition Formelle

$$\alpha_n(t) = \sqrt{n}(F_n(t) - F(t)), \quad t \in \mathbb{R}$$

Représentation Alternative

$$\alpha_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\mathbf{1}_{\{X_i \leq t\}} - F(t)\right)$$

Propriétés de Moments

- **Centrage**: $\mathbb{E}[\alpha_n(t)] = 0 \quad \forall t \in \mathbb{R}$
- **Covariance**:

$$\mathrm{Cov}(\alpha_n(s), \alpha_n(t)) = F(s \wedge t) - F(s)F(t)$$

Propriétés de Structure

- Processus gaussien en limite
- Trajectoires: fonctions en escalier avec sauts aux observations
- Martingale: $\{\alpha_n(t), \mathcal{F}_t\}$ est une martingale

Convergence Ponctuelle Pour chaque t fixé:

$$\alpha_n(t) \xrightarrow{\mathcal{L}} N(0, F(t)(1 - F(t)))$$

Contexte Historique

- **1951**: Donsker introduit l'invariance fonctionnelle
- Problème: Étendre le TCL à des fonctionnelles de F_n
- Innovation: Convergence dans l'espace des fonctions càdlàg

Énoncé Original (Donsker, 1952)

Théorème de Donsker

Le processus empirique $\alpha_n(t) = \sqrt{n}(F_n(t) - F(t))$ converge en loi vers un pont brownien dans l'espace de Skorokhod $D[0, 1]$.

Portée du Résultat

- Premier théorème central limite fonctionnel
- Unifie divers tests d'adéquation
- Fondement de la statistique non-paramétrique moderne

Énoncé Formel Soit $X_1, X_2, \ldots, X_n$ iid de fonction de répartition F continue. Alors:

$$\alpha_n = \sqrt{n}(F_n - F) \xrightarrow{\mathcal{L}} \mathbb{B} \circ F$$

dans l'espace $D[-\infty, +\infty]$ (topologie de Skorokhod), où $\mathbb{B}$ est un pont brownien standard.

Interprétation

- Convergence faible dans un espace fonctionnel
- Pont brownien $\mathbb{B} \circ F$: processus gaussien adapté à F
- Universalité: La limite ne dépend pas de F (si continue)

Conséquence Immédiate Pour toute fonctionnelle ϕ continue:

$$\phi(\alpha_n) \xrightarrow{\mathcal{L}} \phi(\mathbb{B} \circ F)$$

Définition du Pont Brownien

Définition comme Processus Gaussien Un pont brownien $\{\mathbb{B}(t) : t \in [0, 1]\}$ est un processus gaussien centré tel que:

- $\mathbb{B}(0) = \mathbb{B}(1) = 0$
- $\text{Cov}(\mathbb{B}(s), \mathbb{B}(t)) = s \wedge t - st$

Propriétés Fondamentales

- Trajectoires continues (presque sûrement)
- Accroissements corrélés
- Link avec mouvement brownien:

$$\mathbb{B}(t) = W(t) - tW(1)$$

où W est un mouvement brownien standard

Propriétés Statistiques

- $\mathbb{E}[\mathbb{B}(t)] = 0$, $\mathrm{Var}(\mathbb{B}(t)) = t(1-t)$
- $\sup_{t\in[0,1]} |\mathbb{B}(t)|$ a loi de Kolmogorov-Smirnov

Convergence vers le Pont Brownien

Idée de la Preuve

1. **Transformation**: Se ramener au cas F uniforme sur $[0, 1]$
2. **Discrétisation**: Approximation par processus gaussien
3. **Contrôle des fluctuations**: via inégalités maximales
4. **Passage à la limite** dans $D[0, 1]$

Étapes Techniques

- Statistiques d'ordre: $U_{(i)} = F(X_{(i)})$
- Processus empirique uniforme:

$$\mathbb{U}_n(t) = \sqrt{n}(G_n(t) - t)$$

où G_n est la fdr empirique des U_i

Théorème Clé Si F est continue, alors:

$$\alpha_n \xrightarrow{\mathcal{L}} \mathbb{B} \circ F$$

et pour F uniforme:

$$\mathbb{U}_n \xrightarrow{\mathcal{L}} \mathbb{B}$$

Principe d'Invariance

Énoncé

Définition du Principe Si $\phi : D[0,1] \to \mathbb{R}$ est une fonctionnelle continue (topologie de Skorokhod), alors:

$$\phi(\alpha_n) \xrightarrow{\mathcal{L}} \phi(\mathbb{B} \circ F)$$

Conditions de Continuité

- Continuité de Skorokhod: plus faible que continuité uniforme
- Admet des sauts: compatible avec F_n discontinue
- Exemples: norme sup, norme L^2, etc.

Portée Générale Le principe s'étend à:

- Fonctionnelles vectorielles
- Processus indexés

Test de Kolmogorov-Smirnov

- Statistique: $D_n = \sup_{t\in\mathbb{R}} |F_n(t) - F(t)|$
- Fonctionnelle: $\phi(x) = \sup_{t\in[0,1]} |x(t)|$
- Loi limite: $D_n \xrightarrow{\mathcal{L}} \sup_{t\in[0,1]} |\mathbb{B}(t)|$

Test de Cramér-von Mises

- Statistique: $\omega_n^2 = n\int (F_n(t) - F(t))^2 dF(t)$
- Fonctionnelle: $\phi(x) = \int_0^1 x(t)^2 dt$
- Loi limite: $\omega_n^2 \xrightarrow{\mathcal{L}} \int_0^1 \mathbb{B}(t)^2 dt$

Test d'Anderson-Darling

- Statistique: $A_n^2 = n \int \frac{(F_n(t)-F(t))^2}{F(t)(1-F(t))} dF(t)$
- Loi limite: $A_n^2 \xrightarrow{\mathcal{L}} \int_0^1 \frac{\mathbb{B}(t)^2}{t(1-t)} dt$

Classification par Structure

1. **Processus Empirique Usuel (PEU)**: indexé par $\mathbb{R}$
2. **Processus Empirique Indexé (PEI)**: indexé par classes de fonctions
3. **Processus Spécialisés**: quantiles, censure, U-statistiques

Classification par Données

- **IID**: cas standard
- **Dépendantes**: séries temporelles, processus mélangeants
- **Censurées**: données de survie
- **Fonctionnelles**: données courbes

Objectif Commun Étudier la fluctuation:

$$\sqrt{n}(\text{empirique} - \text{théorique})$$

Processus Empirique Usuel (PEU)

Définition

$$\alpha_n(t) = \sqrt{n}(F_n(t) - F(t)), \quad t \in \mathbb{R}$$

Propriétés Clés

- **Convergence**: vers pont brownien $\mathbb{B} \circ F$
- **Covariance**:
  $\text{Cov}(\alpha_n(s), \alpha_n(t)) = F(s \wedge t) - F(s)F(t)$
- **Applications**: tests d'adéquation, intervalles de confiance

Limitations

- Restreint à la fonction de répartition
- Ne capture pas d'autres caractéristiques

Convergence Fonctionnelle

$$\alpha_n \xrightarrow{\mathcal{L}} \mathbb{B} \circ F \quad \text{dans } D(\mathbb{R})$$

Processus Empirique Indexé (PEI)

Définition Générale

$$G_n(f) = \sqrt{n}(P_n f - Pf), \quad f \in \mathcal{F}$$

Classes d'Indexation Typiques

- **Classes d'ensembles**: intervalles, rectangles, boules
- **Classes de fonctions**: Lipschitz, monotones
- **Classes paramétriques**: $\{f_\theta, \theta \in \Theta\}$

Théorème Central Limite Uniforme Sous conditions d'entropie sur $\mathcal{F}$:

$$G_n \xrightarrow{\mathcal{L}} \mathbb{G} \quad \text{dans } \ell^\infty(\mathcal{F})$$

où $\mathbb{G}$ est processus gaussien centré de covariance:

$$\text{Cov}(\mathbb{G}(f), \mathbb{G}(g)) = P(fg) - P(f)P(g)$$

Processus des Quantiles Empiriques

Définition

$$\beta_n(p) = \sqrt{n}(Q_n(p) - Q(p)), \quad p \in [0,1]$$

où $Q_n(p) = \inf\{t : F_n(t) \geq p\}$.
Relation avec PEU Sous régularité ($f = F' > 0$ continue):

$$\beta_n(p) = -\frac{\alpha_n(Q(p))}{f(Q(p))} + o_P(1)$$

Convergence

$$\beta_n \xrightarrow{\mathcal{L}} \frac{\mathbb{B}}{f \circ Q}$$

Processus gaussien de covariance:

$$\mathrm{Cov}(\beta_n(p), \beta_n(q)) = \frac{p \wedge q - pq}{f(Q(p))f(Q(q))}$$

Applications: Intervalles pour quantiles, tests sur la dispersion.

Processus pour Données Censurées à Droite

Cadre

- **Temps de survie**: T_i (non toujours observés)
- **Temps de censure**: C_i
- **Observations**: $Y_i = \min(T_i, C_i)$, $\delta_i = \mathbf{1}_{\{T_i \leq C_i\}}$

Estimateur de Kaplan-Meier

$$\hat{S}_n(t) = \prod_{Y_i \leq t} \left(1 - \frac{\delta_i}{\sum_{j=1}^{n} \mathbf{1}_{\{Y_j \geq Y_i\}}} \right)$$

Processus de Survie Empirique

$$U_n(t) = \sqrt{n}(\hat{S}_n(t) - S(t))$$

Représentation

$$U_n(t) = -S(t) \int_0^t \frac{dM_n(s)}{y(s)} + o_P(1)$$

Convergence Sous censure non-informative:

$$U_n \xrightarrow{\mathcal{L}} U$$

où U est processus gaussien centré.

Covariance Limite

$$\mathrm{Cov}(U(s), U(t)) = S(s)S(t) \int_0^{s \wedge t} \frac{d\Lambda(u)}{y(u)}$$

où $\Lambda =$ fonction de risque cumulative, $y =$ processus à risque.

Applications

- Tests d'égalité de courbes de survie
- Intervalles de confiance pour $S(t)$
- Validation de modèles de survie

Définition

- **Fonction de risque cumulative**: $\Lambda(t) = \int_0^t \lambda(s)ds$
- **Estimateur**:

$$\hat{\Lambda}_n(t) = \sum_{Y_i \leq t} \frac{\delta_i}{\sum_{j=1}^{n} \mathbf{1}_{\{Y_j \geq Y_i\}}}$$

Processus Associé

$$V_n(t) = \sqrt{n}(\hat{\Lambda}_n(t) - \Lambda(t))$$

Propriétés

- **Convergence**: processus gaussien
- **Représentation**:

$$V_n(t) = \sqrt{n} \int_0^t \frac{dM_n(s)}{y(s)} + o_P(1)$$

Applications: Estimation de risque, tests d'hypothèses.

Définition

$$U_n(h) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j)$$

Décomposition de Hoeffding

$$U_n(h) - \theta = \frac{2}{n} \sum_{i=1}^{n} h_1(X_i) + R_n$$

où $h_1(x) = \mathbb{E}[h(x, X)] - \theta$, $R_n = o_P(1/\sqrt{n})$

Processus de U-Statistiques Pour classe $\mathcal{H}$ de noyaux:

$$\{\sqrt{n}(U_n(h) - \theta_h) : h \in \mathcal{H}\}$$

Convergence Si $\mathcal{H}$ Donsker:

$$\sqrt{n}(U_n(h) - \theta_h) \xrightarrow{\mathcal{L}} 2\mathbb{G}(h_1)$$

Processus Empiriques Fonctionnels

Cadre

- **Données fonctionnelles**: $X_1(t), \ldots, X_n(t) \in L^2[0,1]$
- **Moyenne**: $\mu(t) = \mathbb{E}[X(t)]$
- **Moyenne empirique**:

$$\bar{X}_n(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t)$$

Processus Empirique Fonctionnel

$$\alpha_n(t) = \sqrt{n}(\bar{X}_n(t) - \mu(t)), \quad t \in [0,1]$$

Convergence Dans $L^2[0,1]$:

$$\alpha_n \xrightarrow{\mathcal{L}} \mathbb{G}$$

où $\mathbb{G}$ processus gaussien, covariance
$C(s,t) = \mathrm{Cov}(X(s), X(t))$

Processus de Covariance Empirique

Définition

- **Covariance théorique**:
  $C(s,t) = \mathrm{Cov}(X(s), X(t))$
- **Estimateur empirique**:

$$\hat{C}_n(s,t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}_n(s))(X_i(t) - \bar{X}_n(t))$$

Processus de Covariance

$$\mathbb{K}_n(s,t) = \sqrt{n}(\hat{C}_n(s,t) - C(s,t))$$

Convergence Dans $L^2([0,1]^2)$:

$$\mathbb{K}_n \xrightarrow{\mathcal{L}} \mathbb{Z}$$