

UV Statistique pour l'ingénieur

Cours n° 4

Convergence Stochastique

Echantillonnage

- fonction de répartition d'un échantillon
- statistiques d'un échantillon
- loi jointe et vraisemblance

Convergence Stochastique : Rappels

(X_n) = suite de v.a. définies sur un même espace probabilisé

- Cvg en probabilité: si $\exists n_0$ tel que $\forall n > n_0, E(X_n) \rightarrow a$ et $Var(X_n) \rightarrow 0$
alors $(X_n) \xrightarrow{P} a$
- Cvg presque sûre: si $\max_i |X_{n+i} - a| \rightarrow 0$ alors $(X_n) \xrightarrow{PS} a$
- Cvg en loi: $(X_n) \xrightarrow{L} X$ si la suite des fonctions de répartition (F_n) des X_n converge vers la fonction de répartition F de la v.a. X en tout point de continuité
- Propriété : Cvg presque sûre \Rightarrow Cvg en proba. \Rightarrow Cvg en loi

Convergence Stochastique : Exemples

- Les fréquences convergent vers les probabilités.
- La moyenne empirique converge vers l'espérance.
- La loi binomiale $B(n, p)$ converge vers la loi de Poisson $P(\lambda)$
- La loi de Poisson $P(\lambda)$ converge vers la loi de normale $N(\mu, \sigma^2)$

(Ex. de suite de v.a. \Rightarrow estimateur quand la taille de l'échantillon augmente)

Convergence Stochastique : Théorèmes

- Loi des grands nombres :
Si (X_n) une suite i.i.d. de v.a. ayant l'espérance $E(X)$ alors:

$$\frac{(X_1 + \dots + X_n)}{n} \xrightarrow{\text{PS}} E(X)$$

- Théorème central limite :
Si (X_n) une suite i.i.d. de v.a. d'espérance μ et de variance σ^2 alors

$$\frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{L} N(0, 1)$$

$N \rightarrow \infty$

Statistique inférentielle

- Objectif :
 - estimer les statistiques (i.e. moyenne empirique) d'une population à partir d'observations sur un échantillon
- Problème :
 - choisir
 - la taille (fixée en pratique)
 - le type d'échantillonnage
 - Pour assurer la « représentativité » de l'échantillon

Taille:

- $n=10$ petit échantillon,
- $n=100$... moyen moyen
- $n=1000$ grand échantillon

Statistique inférentielle

- Comment assurer la « représentativité » de l'échantillon ?
 - échantillonnage aléatoire simple suite à des tirages équiprobables et indépendants les uns des autres
 - tirage avec remise depuis un ensemble fini ;
 - tirage avec ou sans remise depuis un ensemble infini

échantillon i.i.d.	$= n \text{ v.a. } (X_1, \dots, X_n)$
réalisation de l'échantillon	$= n \text{ observations } (x_1, \dots, x_n)$

- contre-exemples :
 - non indépendance : consommation d'eau
 - non id : temps entre 2 pannes d'un logiciel

Modèle associé à la statistique inférentielle

- 1ère hypothèse :
 - observation x_i = réalisation de la v.a. X (variable parente)
- Modèle :
 - associer à chaque individu i tiré, une v.a. X_i
(pour laquelle on observe une SEULE réalisation x_i)
 \Rightarrow v.a. X_i i.d. (de distribution identique à celle de la v.a. X)
- 2ème hypothèse :
 - X_i mutuellement (ou bien 2 à 2) indépendantes

$\Rightarrow n$ observations : $(x_1, \dots, x_n) = n$ réalisations d'une v.a. X

$\Rightarrow n$ observations : $(x_1, \dots, x_n) =$ réalisation unique du vecteur aléatoire (X, \dots, X) , où X v.a. i.i.d.

A. ROGOZAN – Ph. LERAY

7

Théorie de l'échantillonnage

- Objectif :
 - Étudier (les statistiques d') un échantillon (X_1, \dots, X_n) en fonction de la distribution supposée connue de la variable parente X
 - Étudier (les statistiques) quand la taille de l'échantillon est petite (-), élevée (+) ou moyenne (??)
- Statistique $T =$ v.a. , fonction mesurable de (X_1, \dots, X_n)

$$\Rightarrow T = f(X_1, \dots, X_n)$$

A. ROGOZAN – Ph. LERAY

8

Fonction de répartition *empirique* d'un échantillon

- Définition:

- $F^*(x)$ = proportion des n v.a. X_1, \dots, X_n qui sont inférieures à x
- Pour x_i ordonnées par valeurs croissantes :

$$F^*(x) = \begin{cases} 0 & \text{si } x \leq x_1 \\ \frac{i-1}{n} & \text{si } x_{i-1} < x \leq x_i \\ 1 & \text{si } x > x_n \end{cases}$$

- $\forall x \in \mathfrak{R}, F^*(x)$ est une variable aléatoire,
- F^* une fonction aléatoire (sa réalisation étant une fonction en escalier de \mathfrak{R} dans $[0,1]$ de sauts égaux à $1/n$)

Fonction de répartition *empirique* d'un échantillon

- Propriétés :

- Pour une taille élevée de l'échantillon, $F^*(x)$ tend vers la fonction de répartition de la v.a. X :

$$\forall x \in \mathfrak{R}, F^*(x) \xrightarrow[n \rightarrow \infty]{PS} F(x)$$

- Convergence de $F^*(x)$ vers $F(x)$ presque sûrement uniforme :

$$D_n = \max_x |F^*(x) - F(x)| \text{ suit une loi de probabilité ne dépendant plus de la loi de probabilité de } X$$

Lois des valeurs extrêmes

(pour détecter de valeurs aberrantes dans un échantillon)

- Soient
 - (X_1, \dots, X_n) un échantillon i.i.d
 - $F(x)$ la fonction de répartition de la variable parente X
 - y_1, \dots, y_n les valeurs **ordonnées** des x_1, \dots, x_n
- $\Rightarrow y_1, \dots, y_n =$ réalisation des v.a. (Y_1, \dots, Y_n)

$$Y_n = \underset{i}{\text{Max}} X_i$$

$$H_n(y) = P(Y_n < y) = \prod_{i=1}^n P(X_i < y) = (F(y))^n$$

$$h_n(y) = (nF(y))^{n-1} f(y)$$

$$Y_1 = \underset{i}{\text{Min}} X_i$$

$$H_1(y) = 1 - (1 - F(y))^n$$

$$h_1(y) = n(1 - F(y))^{n-1} f(y)$$

Statistiques d'un échantillon

- Les Indicateurs deviennent des statistiques de l'échantillon :

- Moyenne « empirique » : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \rightarrow \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

- Variance « empirique » : $s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$

- Moments « empiriques »

Statistiques d'un échantillon

- Moyenne **empirique** de l'échantillon : $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- Propriétés :
 - Si X a une espérance $\mu \Rightarrow E(\bar{X}) = \mu$
 - Si X a une variance $\sigma^2 \Rightarrow \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- Lois des Grands Nombres (taille de l'échantillon élevée) :
 - Si X a une espérance $\mu \Rightarrow \bar{X} \xrightarrow[N \rightarrow \infty]{\text{PS}} \mu$
- Application du Th. Central Limite : $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\text{L}} N(0, 1)$

Statistiques d'un échantillon

- Variance **empirique** de l'échantillon : $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
- Lois des grands nombres (taille de l'échantillon élevée) :
 - Si X a une variance $\sigma^2 \Rightarrow S^2 \xrightarrow[N \rightarrow \infty]{\text{PS}} \sigma^2$
- Autre théorème :

$$\frac{S^2 - \frac{(n-1)}{n} \sigma^2}{\sqrt{\text{Var}(S^2)}} \xrightarrow[n \rightarrow \infty]{\text{L}} N(0, 1)$$

$$\text{Var}(S^2) = \frac{n-1}{n^3} [(n-1) \mu_4 - (n-3) \sigma^4]$$

$$\text{Var}(S^2) \simeq \frac{\mu_4 - \sigma^4}{n}$$

Moment centré
d'ordre 4 de X

Statistiques d'un échantillon

- Moment **empirique** non-centré d'ordre k : $\hat{m}_k = \frac{1}{n} \sum_{i=1}^n (X_i)^k$
- Moment **empirique** centré d'ordre k : $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$
 - On retrouve : $\bar{X} = \hat{m}_1$ et $S^2 = \hat{\mu}_2$
- Propriété : Les moments empiriques d'un échantillon i.i.d. (X_1, \dots, X_n) d'une v.a. X convergent p.s. vers les moments théoriques correspondants de X

Exemple : échantillons gaussiens

- Supposons que la v.a. parente $X \sim N(\mu, \sigma^2)$
 - Loi de \bar{X} = combinaison linéaire de n v.a normales

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$
 - Loi de S^2 = loi du **khi-deux** à $n-1$ degrés de libertés

$$n \frac{S^2}{\sigma^2} \sim X_{n-1}^2$$
- \Rightarrow Indépendance entre \bar{X} et S^2

Loi jointe et Vraisemblance

- Soit x_1, \dots, x_n une réalisation d'un échantillon i.i.d. X_1, \dots, X_n (d'une v.a. X)
- Supposons que X suit une loi de probabilité dépendant d'un paramètre réel θ
- Application $\theta \rightarrow L(x_1 \dots x_n; \theta) =$ Vraisemblance du paramètre θ
 - Si X continue : $L(x_1, \dots, x_n; \theta) = f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$
 - Si X discrète : $L(x_1, \dots, x_n; \theta) = P_n(X_1 = x_1, \dots, X_n = x_n; \theta)$

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n P(X_i = x_i; \theta)$$

Loi jointe et Vraisemblance

- Définition de la v.a. $L(x_1, \dots, x_n; \theta)$
 - Probabilité ou densité (à θ fixé) :

$$x_1, \dots, x_n \rightarrow L(x_1, \dots, x_n; \theta)$$
 - Fonction de vraisemblance (à échantillon fixé) :

$$\theta \rightarrow L(x_1, \dots, x_n; \theta)$$