Formulaire de Statistique et Econométrie
Chapitre I : Modes de convergence

Convergence en loi On dit que (X_n) converge en loi vers X ($X_n \xrightarrow{L} X$) si la suite de fonctions (F_n) converge, point par point, vers F :

$$\forall x, \lim_{n \to +\infty} |F_n(X_n) - F(X)| = 0$$

Convergence en probabilité On dit que (X_n) converge en probabilité vers X ($X_n \xrightarrow{P} X$ ou $p\lim_{n\to+\infty} X_n = X$) si

$$\forall \varepsilon > 0, \lim_{n \to +\infty} Pr(|X_n - X| > \varepsilon) = 0$$

La convergence en probabilité est utilisée pour exprimer la loi faible des grands nombres.

Convergence presque sûre On dit que (X_n) converge presque sûrement vers X ($X_n \xrightarrow{ps} X$ si et seulement si

$$\forall \varepsilon > 0, \lim_{n \to +\infty} Pr(|X_i - X| > \varepsilon \ \forall i \geq n) = 0$$

La convergence presque sûre se distingue de la convergence en probabilité par le fait qu'elle est plus générale ($\forall i \geq n$). Intuitivement, elle implique qu'une fois que la série X_n s'approche de X, elle ne s'en écarte pas. On voit rapidement avec la définition (prenons i=n) que la convergence presque sûre implique la convergence en probabilité.

$$X_n \xrightarrow{ps} X \Rightarrow X_n \xrightarrow{P} X$$

La convergence presque sûre est utilisée pour exprimer la loi forte des grands nombres.

Convergence en moyenne quadratique On dit que (X_n) converge en moyenne quadratique vers X ($X_n \xrightarrow{L^2} X$) si

$$\lim_{n \to +\infty} E[||X_n - X||^2] \longrightarrow 0$$

Propriétés : 1. Convergence presque sûre => Convergence en probabilité
2. Convergence en Moyenne Quadratique => Convergence en probabilité
3. $X_n \to X$ en Loi $\Rightarrow \varphi(X_n) \to \varphi(X)$ en loi pour toute fonction *continue* φ
4. $E(X_n) \to c$ et $V(X_n) \to 0 \Rightarrow X_n \to c$ en Moyenne Quadratique

Estimateur convergent : plim $\varphi(\theta_n) = \theta$ (valeur vraie)

Théorème de Slutsky : Si f est une application réelle continue : Xn → X (en probabilité) alors f(Xn) → f(X) (en probabilité)
De même, si g est une application réelle continue : Xn → X (en loi) alors g(Xn) → g(X) (en loi)

Inégalité de Jensen : $g[E(x)] \leq E[g(x)]$ pour g convexe

Rappel :Inégalité de Chebyshev
Soit X une variable aléatoire réelle, discrète ou continue de variance V(X), alors

$$\forall\ \varepsilon > 0, P\Big(|X - E(X)|) > \varepsilon\Big) \leq \frac{V(X)}{\varepsilon^2}$$

Loi faible des grands nombres (Khinchine) Soit (X_i) une suite de variables aléatoires réelles indépendantes et identiquement distribuées (iid) d'espérance finie, $E(X_i) = m$:

$$\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\text{Alors } \bar{X_n} \xrightarrow{p} m$$

Théorème central limite Soit une suite $(X_n)_{n\in\mathbb{N}}$ de variables aléatoires indépendantes, de même loi, d'espérance m et de variance finie σ^2. Notons $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Alors

$$\sqrt{n}\left(\frac{\overline{X}_n - m}{\sigma}\right) \xrightarrow{L} \mathcal{N}(0,1)$$

Statistiques Appliquées
FORMULAIRE chapitre 2 : Echantillonnage

F. GARDES / P. SEVESTRE

1. Echantillonnage sur une population de taille finie :

(a) *Moyenne empirique associée à un échantillon de taille n :*

$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i x_i$ avec $x_i = x(e_i)$ et Z_i le nombre de fois où e_i apparait dans l'échantillon.

- Echantillon tiré avec remise : [Z_i Binomiale]
  $E(\overline{X}_n) = m$ (moyenne vraie dans la population)
  $V(\overline{X}_n) = \frac{\sigma^2}{n}$
- Echantillon tiré sans remise : [lois marginales des $Z_i \sim \mathcal{B}\left(\frac{n}{N}\right)$]
  Proposition 1 : $E(\overline{X}_n) = m$
  Proposition 2 : $V(\overline{X}_n) = \frac{N-n}{N-1} \frac{\sigma_x^2}{n}$
  Proposition 3 : $V(\overline{X}_n) \downarrow$ quand n $\uparrow$ (pour les deux types d'échantillons).

(b) *Variances empiriques associées à un échantillon de taille n :* $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} Z_i \left[x_i - \overline{X}_n\right]^2$

Proposition 4 : $E\left(S_n^2\right) = \sigma_x^2 - V(\overline{X}_n)$
Proposition 5 : (tirage sans remise) $E\left(S_n^2\right) = \frac{n-1}{n} \sigma_x^2$
Proposition 6 : (tirage avec remise) $E\left(S_n^2\right) = \frac{N}{N-1} \frac{n-1}{n} \sigma_x^2$

2. Echantillonnage d'un processus aléatoire (population infinie) : $X_i \sim \mathcal{L}$

(a) *Moyenne empirique :* $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

Proposition 7 : $E(\overline{X}_n) = m$
Proposition 8 : $V(\overline{X}_n) = \frac{\sigma^2}{n}$

(b) *Variance empirique :* $S_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$ et $\overline{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$

Proposition 9 : $E\left(S_n^2\right) = \frac{n-1}{n} \sigma^2$; $E\left(\overline{S}_n^2\right) = \sigma^2$

Proposition 10 : $V(S_n^2)$ et $V(\overline{S}_n^2)$ existent et $\to 0$ quand $n \to +\infty$

3. Echantillons issus d'une loi normale : $X_i \sim \mathcal{N}(m, \sigma^2)$

(a) *Propriétés des moyennes et variances empiriques :*
Proposition 11 : Loi de la moyenne empirique : $\overline{X}_n \sim \mathcal{N}(m, \frac{\sigma^2}{n})$
Proposition 12 : Indépendance entre moyenne empirique et variance empirique : $\overline{X}_n \perp \left(X_i - \overline{X}_n\right)$
Proposition 13 : $\overline{X}_n \perp S_n^2$

(b) *Le Théorème de Fisher :*

i. Les v.a. $(n-1)\, \frac{S_n^2}{\sigma^2}$ et $\frac{\overline{X}_n - m}{\sigma/\sqrt{n}}$ sont indépendantes

ii. $(n-1)\, \frac{S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$

iii. $\frac{X_i - \overline{X}_n}{\sigma} \sim \mathcal{N}\left(0, \sqrt{\frac{n-1}{n}}\right)$

iv. $cov\left(\frac{X_j - \overline{X}_n}{\sigma}\, ;\, \frac{X_k - \overline{X}_n}{\sigma}\right) = -\frac{1}{n} \neq 0$

(c) *Conséquences :*

i. $E\left(S_n^2\right) = \frac{n-1}{n} \sigma^2$ (résultat général).

ii. $V\left(S_n^2\right) = \sigma^4 \frac{2(n-1)}{n^2}$

iii. $E\left(\overline{S}_n^2\right) = \sigma^2$; $V\left(\overline{S}_n^2\right) = 2 \frac{\sigma^4}{n-1}$

iv. $\frac{(\overline{X}_n - m)}{\overline{S}_n/\sqrt{n}} \sim$ *Student* $(n-1)$

Formulaire de Statistique et Econométrie
Chapitre III : Estimation ponctuelle

Estimateur : application T_n de l'échantillon de taille n sur R^n (ou une partie de R^n)

Estimateur sans biais (unbiased) : $E(T_n)= \theta$
Estimateur *asymptotiquement* sans biais : $E(T_n) \to \theta$ quand $n \to \infty$
Estimateur convergent : plim $T_n = \theta$ (valeur vraie)
Estimateur efficace : de variance minimale

Théorème :
1. Tout estimateur sans biais dont la variance tend vers 0 (quand n augmente) est convergent.
2. Tout estimateur *asymptotiquement* sans biais dont la variance tend vers 0 (quand n augmente) est convergent.

Inégalité de Cramer-Rao : Il existe une borne inférieure pour l'ensemble des variances des estimateurs sans biais : $V_\theta(T_n) \geq 1/I_n(\theta)$ avec $I_n(\theta)$ la quantité d'information de Fisher :
$I_n(\theta)=E_\theta(-\partial^2 \ln L/\partial\theta^2) = E_\theta(\partial \ln L/\partial\theta)^2$

Estimation par le principe des *Moindres Carrés* : Min $\sum_{i=1}^n (y_i - \theta)^2$
Estimation par le principe du *Maximum de Vraisemblance* : $\mathrm{Max}_\theta\ L(x_1, x_2,\ldots, x_n ; \theta)$ = Probabilité d'observer le n-uplet $(x_1, x_2,\ldots, x_n)$.
Estimation par la *Méthode des Moments :* résolution des équations : moment empirique de degré $p=f_p(\theta)$

Formulaire de Statistique et Econométrie
Chapitre IV : Estimation par Intervalle de Confiance

Estimation par Intervalle de Confiance d'une *moyenne* X_n :

1. *Grands échantillons :* $X_n \sim N(\mu,\sigma/\sqrt{n})$ par le théorème Central Limite (pour n grand :>30)
2. *Petits échantillons :* - Distribution normale des v.a x_i, σ connue : $X_n \sim N(\mu,\sigma/\sqrt{n})$
   - Distribution normale des v.a x_i, σ inconnue : $\sqrt{n}\,(X_n - mu)/S \sim$ Student à (n-1) degrés de liberté T(n-1), avec S l'écart-type empirique corrigé.

Estimation d'une proportion : $p_n \sim$ loi normale si $np(1-p) \geq 15$; variance = $p(1-p)/n$

v.a. pivotale : fonction du paramètre estimé dont on connaît la distribution statistique ; exemple : $(X_n - \mu)/\ \sigma/\sqrt{n} \sim N(0,1)$

Théorème : Si la distribution est unimodale et symétrique, les Intervalles de Confiance bilatéraux sont *de longueur minimale*.

<u>IC pour les lois usuelles</u> :

(i) Moyenne d'une loi normale : $X_n \pm u_{\alpha/2}\ s'_n/\sqrt{n}$

(ii) Variance d'une loi normale : $(n-1)s'_n/\sigma \sim \chi^2\ (n-1)$
soit $(n-1)s'^2_n/u_{1-\alpha 2} \leq \sigma^2 \leq (n-1)s'^2_n/u_{\alpha 1}$

(iii) Ecart-type d'une loi normale : $s'_n\sqrt{(n-1)/u_{1-\alpha 2}} \leq \sigma \leq s'_n\sqrt{(n-1)/u_{\alpha 1}}$

(iv) IC pour une Proportion : $\sqrt{n}[(X_n - p)/(X_n(1-X_n))^{0.5}]$, donc :
$IC = X_n \pm u_{\alpha/2}\{X_n(1-X_n)/n\}^{0.5}$

(v) IC pour un rapport de variances de lois normales $(n-1)\ s'^2_n/\sigma^2 \sim (n-1) \Rightarrow$
$IC(\sigma^2_B/\sigma^2_A) = [u_{\alpha 1}\ s'^2_B/\ s'^2_A,\ u_{1-\alpha 2}\ s'^2_B/\ s'^2_A]$

(vi) IC pour la différence des espérances de deux lois normales :
- *σ_i connues :* $(X_{1n} - X_{2n}) - (\mu_1 - \mu_2)]/\ \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)} \sim N(0,1)$ (a)
- *σ_i inconnues :* s'^2_i tend en proba vers σ_i^2 , donc IC (a) avec les variances empiriques corrigées s'^2_i.

Formulaire de Statistique et Econométrie
Chapitre VI : Tests d'hypothèses

Erreur de première espèce : Rejet de l'hypothèse nulle H_0 alors qu'elle est vraie
Erreur de seconde espèce : Acceptation de l'hypothèse nulle H_0 alors qu'elle est fausse (H_1 est vraie)

Seuil de signification du test (seuil d'erreur du test): α=probabilité maximale de commettre une erreur de première espèce (par exemple fixée à 5 ou 1%)=$Prob(D_1|H_0)$

Région (dans l'espace des paramètres) *critique* : région de rejet de l'hypothèse nulle H_0
Région d'acceptation : région d'acceptation de l'hypothèse nulle H_0

Test unilatéral : $H_0 : \mu \geq 3$ $H_1 : \mu < 3$
Test bilatéral : $H_0 : \mu = 3$ $H_1 : \mu \neq 3$ (diviser le seuil α entre les deux queues de distribution)

Valeur pivotale : Variable aléatoire dont on connaît la loi, ce qui permet de calculer les paramètres du test. Exemple : $(X_n - \mu)/\sigma/\sqrt{n} \sim N(0,1)$

Tests de moyennes : X_n = moyenne empirique sur un échantillon de taille n

* *σ connu:* - Grands échantillons: $X_n \sim N(\mu, \sigma/\sqrt{n})$ par le théorème Central Limite.

Règle de décision : rejet de H_0 si la moyenne observée X_n n'appartient pas à l'intervalle $[\mu - u_{1-\alpha/2}\ \sigma/\sqrt{n}, \mu + u_{1-\alpha/2}\ \sigma/\sqrt{n}]$ .où $u_{1-\alpha/2}$ désigne le fractile d'ordre $1-\alpha/2$ de la loi N(0,1)

- Petits échantillons (n<30) : $(X_n - \mu)/\ \sigma/\sqrt{n} \sim T(n-1)$

- *σ inconnu:* - Echantillon gaussien: $(X_n - \mu)/s'/\sqrt{n}$ est une valeur pivotale suivant une loi de Student T(n-1) avec s'=écart-type empirique corrigé.
  - Grands échantillons (n>30) : TCL => $(X_n - \mu)/s'/\sqrt{n} \to N(0,1)$ en loi.

Grands échantillons, σ inconnu: prendre l'écart-type empirique corrigé s'=(n/(n-1))s
- Petits échantillons (n<30), σ inconnu: $(X_n - \mu)/s'/\sqrt{n} \sim$ Student à (n-1) degrés de liberté.
- *Cas des proportions :* variance p(1-p)

Tests de variances ou d'écarts-type :

* *Loi normale :* (théorème de Fisher) X_i iid => $(n-1)s'^2/\sigma^2 \sim \chi^2(n-1)$ donc $(n-1)s'^2/\sigma^2$ est une valeur pivotale pour σ.

Règle de décision : rejet de H_0 si s'^2 n'appartient pas à l'intervalle $[u_{\alpha/2}\sigma^2/(n-1), u_{1-\alpha/2}\sigma^2/(n-1)]$

Cas particulier des grands échantillons : $[\sqrt{(n-1)}](s'-\sigma)/\sigma \to N(0,1)$ en loi donc $[\sqrt{(n-1)}]\ (s'-\sigma)/\sigma$ est une valeur pivotale asymptotique pour σ.

Tests de comparaison :

Hypothèse : X_i iid, de loi $L\mu$ dont deux observations de moyennes X_{i1} et X_{i2} sont faites dans 2 échantillons E1 et E2.

$H_0 : \mu_1=\mu_2$ $H_1 : \mu_1\neq\mu_2$ ou $H_0 : \mu_1\leq\mu_2$ $H_1 : \mu_1>\mu_2$

- *Moyennes de variables suivant des lois normales :* X_{i1}~$N(\mu_1,\sigma_1)$ et X_{i2}~$N(\mu_2,\sigma_2)$

σ_i connues : $[X_{1n} - X_{2n} -(\mu_1- \mu_2)]/ \sqrt{}(\sigma_1^2/n_1 + \sigma_2^2/n_2) \sim N(0,1)$ est une valeur pivotable de loi N(0,1)

$\sigma_1= \sigma_2$ inconnues : $[X_{1n} - X_{2n} -(\mu_1- \mu_2)]/ s'\sqrt{}(1/n_1 + 1/n_2) \sim T(n_1+n_2-2)$

$\sigma_1\neq\sigma_2$ inconnues : $[X_{1n} - X_{2n} -(\mu_1- \mu_2)]/ \sqrt{}(s'^2_1/n_1 + s'^2_2/n_2)$ asymtotiquement pivotale pour $(\mu_1- \mu_2)$, tend en loi (si n>30) vers N(0,1).

- *Lois quelquonques :*

Hypothèse : $\sigma_1$et σ_2 connues, grands échantillons (n_i>10)

$(X_{1n} - X_{2n})$ ~$N(\mu_1- \mu_2, (\sigma_1^2/n_1 + \sigma_2^2/n_2)^{0.5})$ asymptotiquement

Hypothèse : $\sigma_1$et σ_2 inconnues, grands échantillons (n_i>50)

$[(X_{1n} - X_{2n})-(\mu_1- \mu_2)]/ (s'^2_1/n_1 + s'^2_2/n_2)^{0.5} \sim N(0,1)$

Test d'indépendance du χ^2 : Caractères X et Y, soit qualitatifs, soit quantitatifs soit respectivement à r et s modalités ; n_{ij} = Card{X=i,Y=j) ; $n = \sum_{ij}n_{ij}$; $p_{ij}= n_{ij}/n$; $p_{i.}$ = probabilité marginale = $\sum_j n_{ij}/n$.

Hypothèse d'indépendance : $H_0 : p_{ij} = p_{i.}\ p_{.j}$

$H_1 : p_{ij} \neq p_{i.}\ p_{.j}$

Statistique : $D_n = \sum_i\sum_j[(n_{ij}-n_{i.}n_{.j}/n)^2/ n_{i.}n_{.j}/n$

$D_n \sim \chi^2[(r-1)(s-1)]$ sous H_0

Tests sur les paramètres d'un modèle linéaire estimés par les MCO : $y_t = \alpha +X_t\beta+\varepsilon_t$

Hypothèse : $\varepsilon \sim N(0, \sigma_\varepsilon)$

- *Test sur un paramètre estimé β^ d'écart-type estimé s:*

  Non rejet de H_0: β=b contre H_1: β ≠ b <=> *(β^-b)/s<u pour une valeur u correspondant sur une loi de Student au seuil de confiance (5 ou 1% : par exemple u=1,697 pour un nombre de degrés de liberté=taille de l'échantillon-nombre de paramètres estimés=30 et un seuil de 5%).*

- *Test sur un vecteur de paramètres estimés β^ de matrice de variances-covariances ∑ :* k=taille du vecteur β^ ; $SCR(M_c)$= somme des carrés des résidus du modèle estimé M.

Test de Fisher comparant un modèle contraint M^c (à (n-p-k) degrés de liberté) par l'hypothèse H_0 au modèle non contraint M^{nc} (estimation libre) à (n-p) degrés de liberté: $\{[SCR(M_c)- SCR(M_{nc})]/k\}/ \{SCR(M_{nc})/(n-p)\}\sim F(k, n-p)$.

Règle de décision : Rejet de H_0 si $\{[SCR(M_c)- SCR(M_{nc})]/k\}/ \{SCR(M_{nc})/(n-p)\}$ > limite de F(k, n-p) pour un seuil de confiance donné (par exemple F(3,60)=2,76 pour un seuil de 5%).