Test de Kolmogorov-Smirnov – Convergence des quantiles empiriques

Soit μ une mesure de probabilité sur $\mathbb{R}$ de fonction de répartition F. On considère une suite de variables aléatoires i.i.d. $(X_n)_{n\geq 0}$ de loi μ.

Définition 1. La mesure empirique d'un échantillon $X_1, \dots, X_n$ est définie par

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

La fonction de répartition empirique d'un échantillon $X_1, \dots, X_n$ est définie par

$$\forall t \in \mathbb{R}, \quad F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}.$$

La mesure μ_n est une mesure de probabilité aléatoire et F_n est sa fonction de répartition. Les résultats ci-dessous montrent qu'un échantillon dénombrable infini caractérise la mesure dont il est issu.

Théorème 2 (Glivenko-Cantelli)**.** *Presque sûrement, la suite des fonctions de répartition empiriques converge uniformément vers la fonction de répartition de μ :*

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| = \|F_n - F\|_\infty \xrightarrow[n\to\infty]{p.s.} 0.$$

La convergence à x fixé découle de la loi des grands nombres appliquée à la suite de variables aléatoires i.i.d. $(\mathbf{1}_{\{X_n\leq x\}})_{n\geq 1}$. Il faut ensuite obtenir la convergence uniforme. On pourra se reporter à [Rev97] pour la démonstration.

Dans le même ordre d'idée, on peut obtenir le résultat suivant, qui reste valable sur $\mathbb{R}^d$.

Théorème 3. *Avec probabilité 1, la suite $(\mu_n)_{n\geq 1}$ converge étroitement vers μ. Plus précisément, presque sûrement, pour toute fonction continue bornée,*

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \int_{\mathbb{R}} f\, d\mu_n \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} f\, d\mu.$$

La convergence presque sure pour une fonction f donnée découle de la loi des grands nombres appliquée à la suite de variables aléatoires i.i.d. intégrables (puisque bornées) $(f(X_i))_{i\geq 1}$. Il reste à permuter « presque sûrement » et « pour toute fonction f ». On pourra se reporter à [Rev97] pour la démonstration.

1 Convergence des quantiles empiriques

On associe à F son inverse généralisée F^- définie sur $[0,1]$ par

$$\forall p \in [0,1], \quad F^-(p) = \inf\{x \in \mathbb{R} \ ; \ F(x) \geq p\}.$$

Si U est une variable aléatoire de loi uniforme sur $[0,1]$ alors $F^-(U)$ est une variable aléatoire de fonction de répartition F. Réciproquement, si F est continue, $F(X)$ suit la loi uniforme sur $[0,1]$ (voir [CGCDM99]). Ceci est faux si F n'est pas continue : si une mesure μ possède un atome alors c'est encore le cas pour toute mesure image de μ.

Définition 4. Un quantile d'ordre p est un réel x tel que $\mathbb{P}(X \leq x) \geq p$ et $\mathbb{P}(X \geq x) \geq 1-p$. Le réel $x_p := F^-(p)$ est un quantile d'ordre p. Pour $p = 1/2$, on parle de médiane.

Pour la loi de Bernoulli de paramètre $1/2$ (resp $3/4$), l'ensemble des médianes est l'intervalle $[0,1]$ (resp. le singleton $\{1\}$).

On suppose dans toute la suite que F est **continue**. Soit $X_1, \ldots, X_n$ un échantillon de loi μ. Presque sûrement, il existe une permutation (aléatoire) σ telle que

$$X_{\sigma(1)} < X_{\sigma(2)} < \cdots < X_{\sigma(n)}.$$

On note $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ l'échantillon réordonné (dans l'ordre croissant). En particulier, $X_{(1)}$ est la plus petite valeur de l'échantillon.

On note $Q_{p,n} = X_{([np]+1)}$ le quantile empirique d'ordre p.

Théorème 5. *Pour tout $p \in]0,1[$, si F possède un unique quantile d'ordre p, qui est alors égal à x_p (c'est-à-dire que F^- est continue en p), alors*

$$Q_{p,n} \xrightarrow[n\to\infty]{p.s.} x_p.$$

Pour tout $p \in]0,1[$, si μ admet une densité f strictement positive au voisinage de x_p, alors

$$\sqrt{n}(Q_{p,n} - x_p) \xrightarrow[n\to\infty]{p.s.} \mathcal{N}(0, \sigma_p^2) \quad \text{où} \quad \sigma_p^2 = \frac{p(1-p)}{f(x_p)^2}.$$

La première partie se déduit du théorème de Glivenko-Cantelli :

$$\begin{aligned} |F(Q_{p,n}) - F(x_p)| &\leq |F(Q_{p,n}) - F_n(Q_{p,n})| + |F_n(Q_{p,n}) - F(x_p)| \\ &\leq \|F_n - F\|_\infty + |([np]+1)/n - p| \xrightarrow[n\to\infty]{p.s.} 0. \end{aligned}$$

On utilise ensuite la continuité de F^- en p pour obtenir le résultat. Pour la convergence en loi, on pourra se référer à [CGCDM99] pour un exemple et [Tas85] pour le résultat.

Exemple 6. On cherche à estimer θ dans les trois modèles suivants :

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} \quad g_\theta(x) = \frac{1}{2} e^{-|x-\theta|} \quad \text{et} \quad h_\theta(x) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}.$$

Comparer les méthodes basées sur la LFGN et le TCL et celle basée sur l'estimation de la médiane empirique.

2 Convergence de la fonction de répartition empirique

Le théorème de Glivenko-Cantelli fournit la convergence presque sûre uniforme de la suite des fonctions de répartition empiriques. Le résultat suivant fournit une vitesse sous une hypothèse supplémentaire..

Théorème 7 (Kolmogorov-Smirnov)**.** *Si F est continue alors*

$$D_n = \sqrt{n}\|F_n - F\|_\infty \xrightarrow[n\to\infty]{\mathcal{L}} \mu_{KS},$$

où μ_{KS} est la loi de Kolmogorov-Smirnov caractérisée par exemple par sa fonction de répartition :

$$\forall t > 0, \quad F_{KS}(t) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2t^2} = 1 + 2\sum_{k=1}^{+\infty} (-1)^k e^{-2k^2t^2}.$$

Remarque 8 (fondamentale)*.* Puisque F est continue et croissante et que F_n est constante par morceaux, le supremum ne peut être atteint qu'au voisinage de points de discontinuité de F_n, c'est-à-dire aux points $(X_i)_{1\le i\le n}$. Plus précisément,

$$\|F_n - F\|_\infty = \max_{1\le i\le n} \left(\left| \frac{i-1}{n} - F(X_{(i)}) \right|, \left| \frac{i}{n} - F(X_{(i)}) \right| \right).$$

De plus, puisque F est continue, $F(X_i)$ suit la loi uniforme sur $[0,1]$ et, comme F est croissante, $(F(X_{(i)}))_{1\le i\le n}$ a même loi qu'un échantillon réordonné de variables aléatoires i.i.d. de loi uniforme sur $[0,1]$. Ainsi, D_n a même loi que

$$\sqrt{n} \max_{1\le i\le n} \left(\left| \frac{i-1}{n} - U_{(i)} \right|, \left| \frac{i}{n} - U_{(i)} \right| \right).$$

La loi de D_n ne dépend donc pas de μ, on dit que c'est une statistique libre.

Il existe de nombreux théorèmes du type de celui de Kolmogorov-Smirnov. Citons celui de Cràmer-Von Mises.

Théorème 9 (Crámer-Von Mises)**.** *Si F est continue alors*

$$nI_n^2 = n\int_{\mathbb{R}} (F_n(x) - F(x))^2 \, dF(x) \xrightarrow[n\to\infty]{\mathcal{L}} \mu_{CVM},$$

où μ_{CVM} est une loi tabulée.

Remarque 10. On peut encore montrer, en découpant l'intégrale sur les intervalles de la forme $[X_{(i)}, X_{(i+1)}]$, que

$$nI_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F(X_{(i)}) \right)^2.$$

C'est encore une statistique libre dont la loi est tabulée pour les petites valeurs de n.

Remarque 11 (Une idée de preuve pour ces théorèmes). La preuve esquissée ci-dessous n'est pas la preuve originale de Kolmogorov mais celle, beaucoup plus élégante, de Doob. On pourra consulter [Fis63] pour de plus amples informations (mais c'est hors programme...).

Pour tout $x \in [0,1]$, on pose $H_n(x) = \sqrt{n}(F_n(x) - x)$. La fonction (aléatoire) H_n est nulle en 0 et 1 et a n sauts de hauteur $1/\sqrt{n}$. En vertu du TCL multidimensionnel, pour $0 \leq x_1 < x_2 < \cdots < x_k < 1$, la suite de vecteurs aléatoires $((H_n(x_1), \ldots, H_n(x_k)))_n$ converge en loi vers un vecteur gaussien centré de matrice de covariance (symétrique) K où $K_{ij} = x_i(1-x_j)$ pour $i \leq j$. On peut en déduire la convergence de la suite de processus aléatoire $(H_n)_n$ vers un processus H appellé pont brownien. il ne reste plus qu'à montrer que le supremum et l'intégrale sont conservés par passage à la limite...

3 Tests associés

On souhaite répondre à la question suivante : les données que j'ai récoltées peuvent-elles être considérées comme distribuées selon la loi $\mathcal{N}(0,1)$? On formalise le problème de la manière suivante. Soit F^0 la fonction de répartition de la loi $\mathcal{N}(0,1)$. On veut tester

$$H_0 \ : \ F = F^0 \quad \text{contre} \quad H_1 \ : \ F \neq F^0.$$

Soit $A_n = \sqrt{n}\left\|F_n - F^0\right\|_\infty$. Sous H_0, $F = F^0$ donc la loi de A_n converge vers la loi de Kolmogorov-Smirnov. Sous H_1, $\left\|F_n - F^0\right\|_\infty$ converge presque sûrement vers $\left\|F - F^0\right\|_\infty > 0$ et A_n converge vers $+\infty$ p.s.

La région de rejet que l'on va choisir sera donc (c'est comme pour le test du χ^2) de la forme $[k_\alpha, +\infty[$ où α est le niveau du test et k_α est le quantile d'ordre $1-\alpha$ de la loi de Kolmogorov-Smirnov.

On peut faire exactement le même raisonnement pour construire un test d'adéquation à partir du thèorème de Crámer-Von Mises.

On peut également tester l'homogénéité de deux distributions : on dispose de deux échantillons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ de fonctions de répartitions empiriques respectives F_n et G_n issus de lois de fonctions de répartition respectives F et G supposées continues. On peut tester

$$H_0 \ : \ F = G \quad \text{contre} \quad H_1 \ : \ F \neq G$$

grâce au résultat suivant :

$$\sqrt{\frac{nm}{n+m}}\|F_n - G_m\| \begin{cases} \xrightarrow[n,m\to\infty]{\mathcal{L}} \mu_{KS} & \text{sous } H_0, \\ \xrightarrow[n,m\to\infty]{p.s.} +\infty & \text{sous } H_1. \end{cases}$$

Références

[CGCDM99] M. Cotrell, V. Genon-Catalot, C. Duhamel et T. Meyre – *Exercices de probabilités*, Cassini, 1999.

[Fis63] M. Fisz – *Probability theory and mathematical statistic*, Wiley, 1963.

[Rev97] D. Revuz – *Probabilités*, Hermann, 1997.

[Tas85] P. Tassi – *Méthodes statistiques*, Économia, 1985.