Cours/TD/TP - Convergence des mesures empiriques Théorème de Glivenko-Cantelli

1 Convergence des mesures empiriques

1.1 Aspect théorique

On suppose donnée une loi de probabilité P inconnue ainsi qu'une suite d'observations $(X_1, \ldots, X_n)$ suivant cette loi. On note alors la mesure empirique associée à ces observations :

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

L'objet de ce paragraphe est de rappeler brièvement les différentes convergences de P_n vers P. On désigne par F_n la fonction de répartition empirique (aléatoire)

$$F_n(t) = P_n(]-\infty; t])$$

Une application directe de la loi des grands nombres assure que $F_n(t)(\omega) \longmapsto F(t)$ p.s. Le théorème de Glivenko-Cantelli donne une convergence *uniforme* de F_n vers F puisque dans la propriété précédente, on a convergence sur un ensemble p.s. qui dépend de t.

Théorème 1.1 (Théorème de Glivenko-Cantelli) *Soit F_n la fonction de répartition empirique d'une fonction de répartition F. Alors*

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \longmapsto 0 \qquad p.s.$$

Remarque 1.2 *Ce théorème a été obtenu en utilisant la loi forte des grands nombres. Une version plus élaborée utilisant alors le théorème central limite aboutira dans un second paragraphe au test de Kolmogorov et Smirnov.*

1.2 Illustration du théorème de de Glivenko-Cantelli

1. Créer un code Matlab permettant d'illustrer le théorème de Glivenko-Cantelli sur un N échantillon de loi binomiale $\mathcal{B}(n,p)$, de loi de Poisson $\mathcal{P}(\lambda)$, de loi exponentielle $\mathcal{E}(\lambda)$ et de loi normale $\mathcal{N}(m, \sigma^2)$.
2. (a) Étant donné un n-échantillon $(X_1, \ldots, X_n)$ de loi de densité f, on note le n-échantillon ré-ordonné par ordre croissant $(X_{(1)}, \ldots, X_{(n)})$. Calculer la densité de la loi de $X_{(1)}$, $X_{(n)}$ et enfin $X_{(i)}$ pour $2 \leq i \leq n$.
   (b) Soient $X_1, \ldots X_{2n-1}$ des v.a. iid de loi μ sur $\mathbb{R}$, de fonction de répartition F et de densité f par rapport à la mesure de Lebesgue. Note m la médiane de μ, $(X_{(1)}, \ldots, X_{(2n-1)})$ les statistiques d'ordre correspondantes. Proposer un estimateur $\hat{m}$ de m convergeant.
   (c) Quelle est la densité de la loi de $\hat{m}$?
   (d) Calculer la densité de $\sqrt{2n-1}(\hat{m} - m)$.
   (e) En déduire que $\sqrt{2n-1}(\hat{m} - m)$ converge en loi vers une mesure gaussienne dont on précisera les paramètres.

2.1 Test d'adéquation à une loi de Kolmogorov-Smirnov

Théorème 2.1 (Kolmogorov-Smirnov) *Soit $(X_1, \dots X_n)$ un échantillon de loi μ sur $\mathbb{R}$ de fonction de répartition F. Si F est continue, alors*

$$K_n = \sqrt{n} \sup_{x \in \mathbb{R}} \|F_n(x) - F(x)\|_\infty \longmapsto^{\mathcal{L}} \mu_{KS}$$

où μ_{KS} est une loi universelle ne dépendant pas de F. Elle est portée par $\mathbb{R}_+$ et a pour fonction de répartition pour $t \geq 0$:

$$F_{KS}(t) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 t^2}$$

1. Montrer que l'expression de K_n s'écrit en réalité :
$$K_n = \sqrt{n} \max_{1 \leq i \leq n} \left\{ \left| F(X_{(i)}) - \frac{i}{n} \right| ; \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$
2. Illustrer le théorème de Kolmogorov-Smirnov qui assure que K_n converge en loi. Pour cela, on pourra tracer un histogramme d'un grand nombre de réalisations de K_n^i pour n assez grand.
3. **Corollaire 2.2** *Soit $(X_1, \dots, X_n)$ un échantillon de loi inconnue μ, de fonction de répartition F. On note F_n sa fonction de répartition empirique associée. On souhaite tester $\mathcal{H}_0 = \acute{} \ \mu = \nu \ \check{}$ contre $\mathcal{H}_1 = \acute{} \ \mu \neq \nu \ \check{}$. Soit $\alpha \in [0;1]$ et $k_{1-\alpha}$ le quantile $1-\alpha$ de la loi de Kolmogorov Smirnov μ_{KS}. Le test qui consiste à rejeter $\mathcal{H}_0$ si $K_n > k_{1-\alpha}$ et à accepter $\mathcal{H}_0$ sinon est asymptotiquement de niveau α, et sa puissance converge vers 1.*
   Créer un code Matlab qui teste le générateur aléatoire uniforme rand au niveau 0.01 avec le test de Kolmogorov-Smirnov. On donne $k_{0.99} = 1.63$.
4. Créer un code Matlab permettant de générer avec l'algorithme de Box-Muller un N échantillon de loi normale $\mathcal{N}(m, \sigma^2)$ affectés par l'utilisateur. Effectuer ensuite un test de Kolmogorov-Smirnov d'adéquation à cette loi. On donne $k_{0.9} = 1.22, k_{0.95} = 1.36$ et $k_{0.99} = 1.63$.
5. On s'intéresse à la durée de vie d'un certain type de matériel et on veut en particulier savoir si cette durée de vie suit une loi exponentielle (hypothèse nulle) ou non (hypothèse $\mathcal{H}_1$). Pour cela, on observe pendant $T = 200$ heures un système où les appareils tombés en panne sont immédiatement replacés. On note N_T le nombre de pannes jusqu'à l'instant T, et par T_i les différents instants de pannes.
   (a) Décrire la loi de N_T sous $\mathcal{H}_0$.
   (b) Sous $\mathcal{H}_0$, décrire la loi de $(T_1, \dots T_n)$ sachant $N_T = n$.
   (c) On a observé $N_T = 5, T_1 = 51, T_2 = 78, T_3 = 110, T_4 = 135$ et $T_5 = 180$. Conclure.

2.2 Test de Cramer von Mises

Ce test est un aménagement du test précédent, sauf que la quantité à regarder n'est plus K_n. En fait, ce test est basé sur la statistique

$$nI_n = n \int (F_n(x) - F(x))^2 \, dF(x)$$

Théorème 2.3 (Cramer von Mises) *Soit $(X_n)_{n \in \mathbb{N}}$ une suite de v.a. i.i.d. de fonction de répartition F continue. On note F_n la fonction de répartition empirique obtenue de l'échantillon de taille n. Alors on a*

$$nI_n \longmapsto^{\mathcal{L}} \sum_{k=1}^{\infty} \frac{Y_k}{\pi k^2}$$

où les v.a. (Y_k) sont i.i.d. de loi $\chi^2(1)$.

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

1. Écrire à nouveau une fonction Matlab testant les générateurs rand et randn grâce aux quantiles $k_{0.99} = 0.74346$ et $k_{0.9} = 0.34730$ dès que $n \geq 100$.
2. On teste la loi exponentielle de paramètre 1, tracer les courbes de puissance en fonction d'un paramètre de perturbation pour comparer les deux tests d'adéquation.
3. Essayer d'autres lois pour voir si les résultats s'inversent.

2.3 Test de comparaison d'échantillons de Kolmogorov-Smirnov

Voici enfin un test qui compare deux échantillons pour savoir s'ils sont issus de mêmes loi ou pas. Il est basé sur le théorème suivant :

Théorème 2.4 (Homogénéité de Kolmogorov-Smirnov) *Soit $(X_1, \ldots, X_n)$ un n-échantillon de fonction de répartition F, et $(Y_1, \ldots, Y_m)$ un m-échantillon de fonction de répartition G. On suppose que ces deux échantillons sont indépendants et que F et G sont continues. On veut tester $\mathcal{H}_0 = \text{´ } F = G \text{ ˇ}$ contre $\mathcal{H}_1 = \text{´ } F \neq G \text{ ˇ}$. On note F_n et G_m les fonctions de répartition empirique obtenues, alors sous $\mathcal{H}_0$, on a*

$$\sqrt{\frac{nm}{n+m}} \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)| \longmapsto^{\mathcal{L}} \mu_{KS}$$

1. Effectuer un test d'homogénéité de Kolmogorov-Smirnov sur deux échantillons indépendants de loi uniforme $\mathcal{U}([0;1])$ avec $n = 100$ et $m = 1000$.
2. On a relevé dans deux forêts les hautes en mètres de 12 et 14 arbres respectivement. On obtient le tableau suivant :

Foret 1	23.4	24.4	24.6	24.9	25	26.2	26.3	26.8	26.9	27	27.6	27.7	
Foret 2	22.5	22.9	23.7	24	24.4	24.5	25.3	26	26.2	26.4	26.7	26.9	27.4	28.5

Tester l'homogénéité des deux forêts.