L3 Mapi3 Simulations Stochastiques (stats)

Clément Pellegrini

clement.pellegrini@math.univ-toulouse.fr

Institut de Mathématiques de Toulouse,
Equipe de Statistique et Probabilité,
Bureau 220 Bâtiment 1R1

Chapitre 2) Test non-paramétriques

Introduction

- On va aborder des test dits "non paramétrique" c'est à dire que l'on va tester le jeu d'hypothèse

$$\begin{cases} H_0 : & F = F_{ref} \\ H_1 : & F \neq F_{ref} \end{cases}$$

où F_{ref} correspond à une fonction de répartition (f.d.r) de référence et on a un échantillon

$$X_1, \ldots, X_n \sim F$$

où F est une f.d.r inconnue et les X_i sont i.i.d

- On parle de test non paramétriques car F est décrite par un nombre infini de paramètres $(F(t), t \in \mathbb{R})$
- On rappelle que $F(t) = \mathbb{P}[X_1 \leqslant t] = \mathbb{P}[X_i \leqslant t]$

Introduction

- On va aborder des test dits "non paramétrique" c'est à dire que l'on va tester le jeu d'hypothèse

$$\begin{cases} H_0 : & F = F_{ref} \\ H_1 : & F \neq F_{ref} \end{cases}$$

où F_{ref} correspond à une fonction de répartition (f.d.r) de référence et on a un échantillon

$$X_1, \ldots, X_n \sim F$$

où F est une f.d.r inconnue et les X_i sont i.i.d

- On parle de test non paramétriques car F est décrite par un nombre infini de paramètres $(F(t), t \in \mathbb{R})$
- On rappelle que $F(t) = \mathbb{P}[X_1 \leqslant t] = \mathbb{P}[X_i \leqslant t]$

Fonction de répartition empirique

- L'objet mathématique qui va nous permettre de faire l'estimation de la fonction F inconnue est la fonction de répartition empirique

$$\begin{array}{rccc} F_n : & \mathbb{R} & \to & [0,1] \\ & t & \to & \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \end{array}$$

Fonction de répartition empirique

- L'objet mathématique qui va nous permettre de faire l'estimation de la fonction F inconnue est la fonction de répartition empirique

$$\begin{array}{rccc} F_n : & \mathbb{R} & \to & [0,1] \\ & t & \to & \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \end{array}$$

Fonction de répartition empirique

- La valeur

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

dépend des valeurs prises par l'échantillon $X_1, \ldots, X_n$ qui est aléatoire donc la fonction F_n est une fonction aléatoire.

- Si on définit la mesure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

alors

$$F_n(t) = \mu_n(]-\infty, t])$$

- C'est donc bien une f.d.r, on l'appelle empirique (comme pour la moyenne empirique) car elle dépend des résultats de l'expérience.
- Dessin

Fonction de répartition empirique

- La valeur

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

dépend des valeurs prises par l'échantillon $X_1, \ldots, X_n$ qui est aléatoire donc la fonction F_n est une fonction aléatoire.

- Si on définit la mesure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

alors

$$F_n(t) = \mu_n(]-\infty, t])$$

- C'est donc bien une f.d.r, on l'appelle empirique (comme pour la moyenne empirique) car elle dépend des résultats de l'expérience.
- Dessin

Fonction de répartition empirique

- La valeur

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

dépend des valeurs prises par l'échantillon $X_1, \ldots, X_n$ qui est aléatoire donc la fonction F_n est une fonction aléatoire.

- Si on définit la mesure

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}$$

alors

$$F_n(t) = \mu_n(]-\infty, t])$$

- C'est donc bien une f.d.r, on l'appelle empirique (comme pour la moyenne empirique) car elle dépend des résultats de l'expérience.
- Dessin

Fonction de répartition empirique

- Fixons $t \in \mathbb{R}$ comme les v.a (X_i) sont i.i.d alors les v.a

$$(\mathbf{1}_{X_i \leqslant t})$$

sont également i.i.d et on a que pour tout i

$$\mathbb{E}(\mathbf{1}_{X_i \leqslant t}) = \mathbb{P}[X_i \leqslant t] = \mathbb{P}[X_1 \leqslant t]$$

- On peut donc appliquer la LFGN

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \xrightarrow[n \to +\infty]{\text{p.s.}} \mathbb{P}[X_1 \leqslant t] = F(t)$$

Fonction de répartition empirique

- Fixons $t \in \mathbb{R}$ comme les v.a (X_i) sont i.i.d alors les v.a

$$(\mathbf{1}_{X_i \leqslant t})$$

sont également i.i.d et on a que pour tout i

$$\mathbb{E}(\mathbf{1}_{X_i \leqslant t}) = \mathbb{P}[X_i \leqslant t] = \mathbb{P}[X_1 \leqslant t]$$

- On peut donc appliquer la LFGN

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \xrightarrow[n \to +\infty]{\text{p.s.}} \mathbb{P}[X_1 \leqslant t] = F(t)$$

Fonction de répartition empirique

- Les v.a $(\mathbf{1}_{X_i \leqslant t})$ sont des v.a i.i.d qui suivent la loi $\mathcal{B}(\mathbb{P}[X_1 \leqslant t]) = \mathcal{B}(F(t))$
- La v.a $nF_n(t)$ compte donc le nombre de succés ainsi

$$nF_n(t) \sim \mathcal{B}(n, F(t))$$

- On a appliqué la LFGN, le TCL nous donne

$$\sqrt{n}\left(F_n(t) - F(t)\right) \xrightarrow[n\to+\infty]{\text{Loi}} \mathcal{N}(0, F(t)(1-F(t))$$

$$\frac{\sqrt{n}}{\sqrt{F(t)(1-F(t))}}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}_{X_i \leqslant t} - F(t))\right) \xrightarrow[n\to+\infty]{\text{Loi}} \mathcal{N}(0,1)$$

Fonction de répartition empirique

- Les v.a $(\mathbf{1}_{X_i \leqslant t})$ sont des v.a i.i.d qui suivent la loi $\mathcal{B}(\mathbb{P}[X_1 \leqslant t]) = \mathcal{B}(F(t))$
- La v.a $nF_n(t)$ compte donc le nombre de succés ainsi

$$nF_n(t) \sim \mathcal{B}(n, F(t))$$

- On a appliqué la LFGN, le TCL nous donne

$$\sqrt{n}\left(F_n(t) - F(t)\right) \xrightarrow[n\to+\infty]{\text{Loi}} \mathcal{N}(0, F(t)(1-F(t))$$

$$\frac{\sqrt{n}}{\sqrt{F(t)(1-F(t))}}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{1}_{X_i \leqslant t} - F(t))\right) \xrightarrow[n\to+\infty]{\text{Loi}} \mathcal{N}(0,1)$$

Fonction de répartition empirique

- Jusqu'ici on a considéré des résultats lorsque t était fixé et on a obtenu des convergences lorsque F_n était évalué en t.
- Si on fixe un t_0 alors pour un n_0 grand on peut dire des choses sur la proximité de $F_n(t_0)$ et $F(t_0)$ mais si on change et qu'on regarde un temps t_1 alors on aura un autre n_1 qui n'a aucun lien à priori avec le premier n_0.
- Ainsi à chaque $t \in \mathbb{R}$ correspond un n_t suffisamment grand tel que $F_n(t)$ est proche de $F(t)$. Ce résultat n'est pas satisfaisant car on voudrait pouvoir prendre un n suffisament grand indépendant de t. Une sorte d'uniformité.
- On aimerait établir des résultats du type

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

Fonction de répartition empirique

- Jusqu'ici on a considéré des résultats lorsque t était fixé et on a obtenu des convergences lorsque F_n était évalué en t.
- Si on fixe un t_0 alors pour un n_0 grand on peut dire des choses sur la proximité de $F_n(t_0)$ et $F(t_0)$ mais si on change et qu'on regarde un temps t_1 alors on aura un autre n_1 qui n'a aucun lien à priori avec le premier n_0.
- Ainsi à chaque $t \in \mathbb{R}$ correspond un n_t suffisamment grand tel que $F_n(t)$ est proche de $F(t)$. Ce résultat n'est pas satisfaisant car on voudrait pouvoir prendre un n suffisament grand indépendant de t. Une sorte d'uniformité.
- On aimerait établir des résultats du type

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

Glivenko-Cantelli

- On a le résultat suivant appelé parfoirs théorème fondamental de la statistique ou théorème de Glivenko Cantelli

Theorem (Glivenko Cantelli)

Soit (X_n) *une suite de v.a.i.i.d de f.d.r F alors*

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

- On doit donc passer d'une convergence simple à une convergence uniforme

Glivenko-Cantelli

- Maintenant qu'on a un résultat de convergence presque sûre uniforme de F_n vert F on a donc un résultat de consistence. On voudrait maintenant pouvoir faire le test.
- Pour cela nous avons le résultat suivant

Theorem (Kolmogorov)

Soit (X_n) *une suite de v.a.i.i.d de f.d.r F CONTINUE alors la loi de la v.a*

$$D_n = \|F_n - F\|_\infty$$

ne dépend pas de F

- Notons qu'il y a une contrainte c'est la continuité de F.

- Revenons au Théorème de Kolmogorov

Theorem (Kolmogorov)

Soit (X_n) une suite de v.a.i.i.d de f.d.r F CONTINUE alors la loi de la v.a

$$D_n = \|F_n - F\|_\infty$$

ne dépend pas de F

Utilisation

- On a donc besoin de connaître les quantiles de la loi

$$D_n = \|F_n - F\|_\infty$$

- On définit $d_{n,1-\alpha}$ tel que

$$\alpha = \mathbb{P}[D_n > d_{n,1-\alpha}]$$

- Ces valeurs sont connues (des bonnes approximations) tabulées.

Utilisation

- On s'intérresse souvent à D_n pour des valeurs grandes. En particulier il existe W_{max} telle que

$$\sqrt{n}D_n \xrightarrow[n\to+\infty]{\text{Loi}} W_{max}$$

- En particulier pour tout $x \geqslant 0$ et n grand

$$\mathbb{P}[\sqrt{n}D_n \leqslant x] \simeq \mathbb{P}[W_{max} \leqslant x]$$

- On en déduit que

$$\mathbb{P}[D_n > \frac{x}{\sqrt{n}}] \simeq \mathbb{P}[W_{max} > x]$$

et donc

$$d_{n,1-\alpha} \simeq \frac{w_{1-\alpha}}{\sqrt{n}}$$

- Pour n grand les quantiles de W_{max} donnent ceux de D_n

Utilisation

- On s'intérresse souvent à D_n pour des valeurs grandes. En particulier il existe W_{max} telle que

$$\sqrt{n}D_n \xrightarrow[n\to+\infty]{\text{Loi}} W_{max}$$

- En particulier pour tout $x \geqslant 0$ et n grand

$$\mathbb{P}[\sqrt{n}D_n \leqslant x] \simeq \mathbb{P}[W_{max} \leqslant x]$$

- On en déduit que

$$\mathbb{P}[D_n > \frac{x}{\sqrt{n}}] \simeq \mathbb{P}[W_{max} > x]$$

et donc

$$d_{n,1-\alpha} \simeq \frac{w_{1-\alpha}}{\sqrt{n}}$$

- Pour n grand les quantiles de W_{max} donnent ceux de D_n

Utilisation

- On est en mesure de présenter le test de Kolmogorov qui va nous permettre de tester

$$\begin{cases} H_0 : & F = F_{ref} \\ H_1 : & F \neq F_{ref} \end{cases}$$

- On pose le test

$$\varphi_n(X_1, \ldots, X_n) = \mathbf{1}_{D_n > d_{n,1-\alpha}},$$

avec $D_n = \|F_n - F_{ref}\|_\infty$

- Le résultat reste vrai même si F_{ref} n'est pas continue.
- On montrera en T.D que ce test est consistent i.e sous H_1

$$\mathbb{P}[\varphi_n(X_1, \ldots, X_n) = 1] \xrightarrow[n \to +\infty]{} 1$$

Utilisation

- On est en mesure de présenter le test de Kolmogorov qui va nous permettre de tester

$$\begin{cases} H_0 : & F = F_{ref} \\ H_1 : & F \neq F_{ref} \end{cases}$$

- On pose le test

$$\varphi_n(X_1, \ldots, X_n) = \mathbf{1}_{D_n > d_{n,1-\alpha}},$$

avec $D_n = \|F_n - F_{ref}\|_\infty$

- Le résultat reste vrai même si F_{ref} n'est pas continue.
- On montrera en T.D que ce test est consistent i.e sous H_1

$$\mathbb{P}[\varphi_n(X_1, \ldots, X_n) = 1] \xrightarrow[n \to +\infty]{} 1$$

Soit $(X_1, \ldots, X_n)$ un n-échantillon de fonction de répartition F inconnue. Notre seule connaissance sur F est qu'elle est continue. On note $\mathbf{F}_n$ la fonction de répartition empirique associée Ã l'échantillon :

$$\forall t \in \mathbf{R}, \qquad \mathbf{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leqslant t} \, .$$

On notera F^{-1} l'inverse généralisée de la fonction F.

1. On rappelle que si $U_1, \ldots, U_n$ sont des variables aléatoires i.i.d. uniformes sur $[0,1]$, alors $\left(F^{-1}(U_1), \ldots, F^{-1}(U_n)\right)$ est un n-échantillon de fonction de répartition F. En déduire que la loi de $\|\mathbf{F}_n - F\|_\infty$ est la mÃªme quelle que soit F continue.
   Dans la suite, pour tout $\alpha \in]0,1[$, on notera $d_{n,1-\alpha}$ un quantile d'ordre $1-\alpha$ de cette loi.
2. En cours, nous avions construit un intervalle de confiance centré en la moyenne empirique qui contenait le vrai paramètre θ d'une loi de Bernoulli avec probabilité supérieure ou égale Ã $1-\alpha$. En vous inspirant de ce résultat, construisez une région de confiance $\widehat{\mathcal{F}_n} \subset [0,1]^{\mathbf{R}}$ qui contienne la vraie fonction de répartition F avec probabilité supérieure ou égale Ã $1-\alpha$.

1. Application : vous observez les valeurs d'échantillon suivantes :

   0.38 0.16 0.04 0.63 0.44 0.27 0.51 0.32 0.06 0.13

2. Représentez graphiquement la fonction $t \mapsto \mathbf{F}_n(t;\omega)$ sur cet exemple. Expliquez brièvement comment on pourrait représenter sur le graphique la région de confiance $\widehat{\mathcal{F}_n}(\omega)$, pour le niveau de risque $\alpha = 5\%$.
   On donne quelques valeurs approchées du quantile $d_{10,1-\alpha}$:

α	0.01	0.025	0.05	0.1
$d_{10,1-\alpha}$	0.489	0.445	0.409	0.369

3. Que signifie très concrètement la valeur 5% ?