TD 9 - jeudi

jeudi 8 avril 2021 08:32

Exercice 1

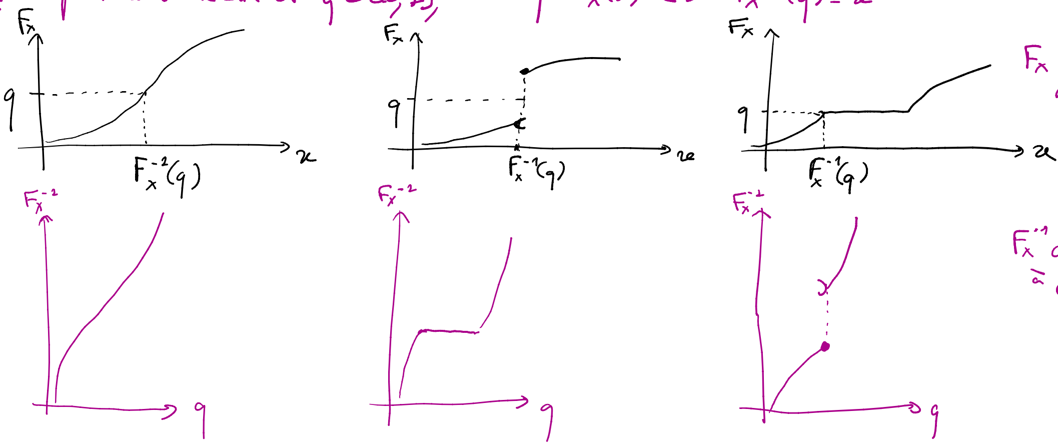
1) $F_X : x \in \mathbb{R} \longmapsto \mathbb{P}(X \leq x) \in [0,1]$. Croissante, $\lim_{-\infty} F_X = 0$, $\lim_{+\infty} F_X = 1$.

Si X est une variable aléatoire qui vaut 0 p.s., alors $F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$

ce qui n'est ni bijectif, ni injectif, ni surjectif.

2) Pour tout $q \in [0,1]$, $F_X^{-1}(q) = \inf\{x \in \mathbb{R} : F_X(x) \geq q\}$.

Propriété utile : pour tout $x \in \mathbb{R}$ et $q \in [0,1]$, $q \leq F_X(x) \iff F_X^{-1}(q) \leq x$



F_X continue à droite

F_X^{-1} continue à gauche

3) NON, aucune des deux égalités n'est vraie en général car cela voudrait dire que F_X est toujours bijective, et on a donné un contre-exemple.

a) D'après la propriété, $\forall x \in \mathbb{R}$, $F_X(x) \leq F_X(x)$ donc $\boxed{F_X^{-1}(F_X(x)) \leq x}$.

b) De même, $\forall q \in [0,1]$, $F_X^{-1}(q) \leq F_X^{-1}(q)$ donc $\boxed{q \leq F_X(F_X^{-1}(q))}$

4) On utilise $F_X^{-1}(U) \leq x \iff U \leq F_X(x)$, donc

$$\mathbb{P}(F_X^{-1}(U) \leq x) = \underbrace{\mathbb{P}(U \leq F_X(x))}_{\mathbb{P}(U \in [0, F_X(x)])} \overset{\text{car } U \sim \mathcal{U}([0,1])}{=} F_X(x) \underset{\text{déf. de } F_X}{=} \mathbb{P}(X \leq x)$$

(Si $U \sim \mathcal{U}([a,b])$ et $[c,d] \subset [a,b]$, alors $\mathbb{P}(U \in [c,d]) = \frac{d-c}{b-a}$)

Les fonctions de répartition de $F_X^{-1}(U)$ et de X sont les mêmes, donc ces variables ont la même loi.

5) « La loi de X n'a pas d'atome » : on a donc F_X continue, F_X est surjective $\mathbb{R} \to \underline{\mathrm{Im}(F_X)}$

$(0,1), [0,1), (0,1], [0,1]$

donc pour tout $q \in (0,1)$, il existe $y \in \mathbb{R}$ tel que $F_X(y) = q$.

Soit $q \in (0,1)$, soit $z \in \mathbb{R}$ le plus grand réel tel que $F_X(z) = q$. $F_X(z) \leq q$

$\mathbb{P}(F_X(X) \leq q) = ?$

(But : $= q$)

(car alors $= \mathbb{P}(U \leq q)$ où $U \sim \mathcal{U}([0,1])$)

$F_X(X)$ → en quel point est évalué F_X ; quelle est la variable donc on considère la fonction de répartition

$F_X(X) = \mathbb{P}(X' \leq x | X)$ où X' et X ont la même loi.

X est une v.a. à valeurs dans $\mathbb{R}$.

Si $f : \mathbb{R} \to [0,1]$ on peut définir $f(X)$ une v.a. à valeurs dans $[0,1]$

quelle est ... donc on considère la fonction de répartition

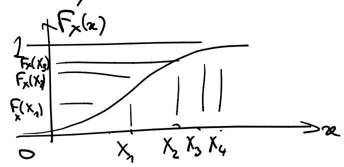
Si $f : \mathbb{R} \to [0,1]$, on peut définir $f(X)$ une v.a. à valeurs dans $[0,1]$

Prenez $f = \boxed{F_X : t \mapsto \mathbb{P}(X \leq t)}$

Soit X' de même loi que X

$F_{X'}(X)$, $F_X(X')$, $F_X(X)$, $F_{X'}(X')$ ont la même loi.

$\mathbb{P}(X \leq X) = 1$, alors que



$$\mathbb{P}(F_X(X) \leq q) = \mathbb{P}(F_X(X) \leq F_X(z)) = \mathbb{P}(X \leq z) = F_X(z) = q.$$

(car $F_X(y) \leq F_X(z) \iff y \leq z$ par définition de z)

Exercice 2

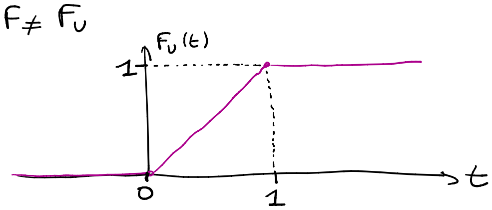
On note $X_1, \dots, X_n$ les observations (avec ici $n = 20$).
On suppose les X_i iid de loi $\mathbb{P}$.

1) $H0 : \mathbb{P} = \mathcal{U}([0,1])$ $\quad H1 : \mathbb{P} \neq \mathcal{U}([0,1])$

Si F est la fonction de répartition de X_1 et F_U la fonction de répartition de la loi $\mathcal{U}([0,1])$,

$H0 : F = F_U$ $\quad H1 : F \neq F_U$

2) $F_U : t \in \mathbb{R} \mapsto \begin{cases} 0 & \text{si } t \leq 0 \\ t & \text{si } t \in (0,1) \\ 1 & \text{si } t \geq 1 \end{cases}$



3) Ne pas confondre $X_1, \dots, X_n$ et $\underbrace{X_{(1)}, \dots, X_{(n)}}$

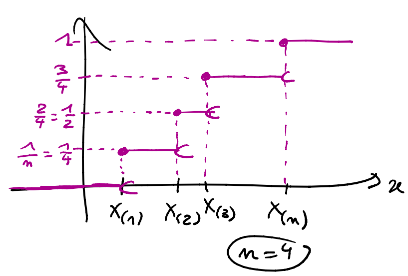
statistiques d'ordre : c'est l'échantillon $X_1, \dots, X_n$ trié par ordre croissant, $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$

Pour tout $x \in \mathbb{R}$,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \leq x} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_{(i)} \leq x}$$

$$= \begin{cases} 0 & \text{si } x < X_{(1)} \\ \frac{1}{n} \max\{i \in \{1, \dots, n\} : X_{(i)} \leq x\} \end{cases}$$

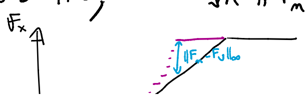
$$= \begin{cases} 0 & \text{si } x < X_{(1)} \\ \frac{i}{n} & \text{si } x \in [X_{(i)}, X_{(i+1)}) \\ 1 & \text{si } x \geq X_{(n)} \end{cases}$$



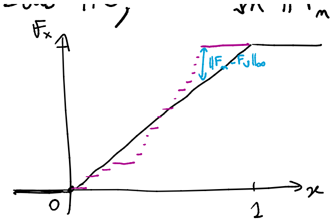
4) Lorsque les X_i ont pour fonction de répartition F,

$\|F_n - F\|_\infty \xrightarrow[n \to +\infty]{} 0$. (Sous $H0$, $\|F_n - F_U\|_\infty \xrightarrow[n \to +\infty]{} 0$)

5) Sous $H0$, $\sqrt{n} \|F_n - F_U\|_\infty \xrightarrow[n \to +\infty]{} \underbrace{\sup_{t \in [0,1]} |B(t)|}_{\text{suit la loi de Kolmogorov.}}$ où B est le pont brownien sur $[0,1]$.



5) Sous H_0, $\sqrt{n}\,\|F_n - F_U\|_\infty \xrightarrow[n\to+\infty]{} \underbrace{\sup_{t\in[0,1]} |B(t)|}_{\text{suit la loi de Kolmogorov.}}$ où B est le pont brownien sur $[0,1]$.



me dépend pas de la loi des X_i !

6) D'après le théorème de Glivenko-Cantelli, $\|F_n - F^*\|_\infty \xrightarrow[n\to+\infty]{} 0$

Si $X_1, \dots, X_n$ sont iid de fonction de répartition F^*, et F_n est la fn de répartition empirique, alors

« $F_n \to F^*$ en norme infinie »

$$\Big| \|F_n - F_U\|_\infty - \|F^* - F_U\|_\infty \Big| \leq \|F_n - F^*\|_\infty \text{ par l'inégalité triangulaire}$$

$\big| |a-b| - |a-c| \big| \leq |b-c|$

$a = F_U(t)$, $b = F_n(t)$, $c = F^*(t)$

puis $\sup_{t\in[0,1]}$

donc $\boxed{\|F_n - F_U\|_\infty \xrightarrow[n\to+\infty]{} \|F^* - F_U\|_\infty}$.

$\|F_n - F_U\|_\infty \xrightarrow{?} \|F^* - F_U\|_\infty$, $F_n \to F^*$

Ainsi, $\boxed{\sqrt{n}\,\|F_n - F_U\|_\infty \xrightarrow[n\to+\infty]{} +\infty}$ lorsque $F^* \neq F_U$

$\|F_n - F_U\|_\infty - \|F^* - F_U\|_\infty \xrightarrow[n\to+\infty]{} 0$

Puis $\sqrt{n}\,\|F_n - F_U\|_\infty \underset{n\to+\infty}{\sim} \sqrt{n}\,\underbrace{\|F^* - F_U\|_\infty}_{\text{constante} > 0} \xrightarrow[+\infty]{} +\infty$

7) On prend comme statistique de test $\sqrt{n}\,\|F_n - F_U\|_\infty$.

Cette quantité tend vers $+\infty$ sous H_1, d'où une région de rejet de la forme $R = [t, +\infty[$.

On prend t le quantile d'ordre 95% de la loi de Kolmogorov.

8) Le tableau A.14, version « Two-Sided », contient les quantiles de $\|F_n - F_U\|_\infty$ pour différentes valeurs de n.

Two-Sided : on prend comme statistique $\sup_{t\in[0,1]} |F_n(t) - F_U(t)| = \max\Big(\sup_t (F_n(t) - F_U(t)), \sup_t (F_U(t) - F_n(t)) \Big)$

One-Sided : $\sup_{t\in[0,1]} (F_n(t) - F_U(t))$

↳ pour tester $F > F_U$, $F < F_U$ …

On prend donc $\boxed{t = \sqrt{n} \times 0{,}294}$.

(Exercice 3 : tableau A.16)

9) Le maximum est atteint en $X_{(2)}$.

$$\|F_n - F_U\|_\infty = \max_{1\leq i\leq n} \left(\max\left(F_U(X_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F_U(X_{(i)}) \right) \right)$$

→ Prenez $i = 2$.

$\|F_n - F_U\| = 0{,}402$. $\overset{i=2}{=} \max\Big(\underbrace{\overbrace{0{,}452 - \tfrac{1}{20}}^{\boxed{0{,}402}}}_{F_U(X_{(2)})}, \overbrace{\tfrac{2}{20} - 0{,}452}^{-0{,}352} \Big)$

10) $T = \sqrt{n}\,\|F_n - F_U\|_\infty$, $T^{obs} = \sqrt{n} \times 0{,}402$.

$\sqrt{n} \times 0{,}402 \in [\sqrt{n} \times 0{,}294 ; +\infty[$ donc on rejette $H0$: les X_i ne suivent pas la loi $\mathcal{U}([0,1])$.

11)

$[0, \frac{1}{3}[$	$[\frac{1}{3}, \frac{2}{3}[$	$[\frac{2}{3}, 1]$
$1 = N_1$	$10 = N_2$	$9 = N_3$

$(N_1, N_2, N_3) \sim \mathcal{M}ult(20, p)$ avec p inconnu.

Test de $H0 : p = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ vs. $H1 : p \neq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

Suite du test : voir TD 8.

Exercice 3 Reprenez la démarche du cours !

Notons $X_1, \dots, X_n$ le premier échantillon et $Y_1, \dots, Y_m$ le second ($n = 5$ et $m = 10$).

Notons $h(X,Y) = \| F_n - G_m \|_\infty$

F_n : fonction de répartition empirique de $(X_1, \dots, X_n)$

G_m : fonction de répartition empirique de $(Y_1, \dots, Y_m)$

Notons F_X la fonction de répartition (inconnue !) de X et F_Y celle de Y.

$H_0 : F_X = F_Y$

$H_1 : F_X \neq F_Y$

$$h(X,Y) = \sup_{x \in \mathbb{R}} \Big| \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le x}}_{F_n(x)} - \underbrace{\frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{Y_j \le x}}_{G_m(x)} \Big|$$

U_i, V_j des v.a. i.i.d de loi $\mathcal{U}([0,1])$

$$\sim \sup_{x \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{F_X^{-1}(U_i) \le x} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{F_Y^{-1}(V_j) \le x} \Big|$$

$$\overset{\text{sous } H_0}{\sim} \sup_{x \in \mathbb{R}} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{F_X^{-1}(U_i) \le x} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{F_X^{-1}(V_j) \le x} \Big|$$

$$h(X,Y) \sim \sup_{q \in \mathrm{Im}(F_X)} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \le q} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{V_j \le q} \Big|$$

et la loi du terme de droite ne dépend ni de F_X ni de F_Y, seulement de n et m !

en fait si via le « $q \in \mathrm{Im}(F_X)$ »
mais en tout cas
$$h(X,Y) \le \sup_{q \in (0,1)} | \dots |$$

Notons $d_{n,m,\alpha}$ le quantile de niveau α de la v.a. $\displaystyle \sup_{q \in (0,1)} \Big| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \le q} - \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{V_j \le q} \Big|$.

Alors $\mathbb{P}(h(X,Y) \le d_{n,m,\alpha}) \ge \alpha$

Sous H_1, $h(X,Y) \xrightarrow[m \to +\infty]{n \to +\infty} \underbrace{\| F_X - F_Y \|_\infty}_{>0}$ alors que sous H_0, $h(X,Y) \xrightarrow[m \to +\infty]{n \to +\infty} 0$

(car $h(X,Y) = \| F_n - G_m \|_\infty$
$\le \underbrace{\| F_n - F_X \|_\infty}_{\to 0} + \underbrace{\| F_X - F_Y \|_\infty}_{= 0 \text{ sous } H_0} + \underbrace{\| F_Y - G_m \|_\infty}_{\to 0}$)
Glivenko-Cantelli

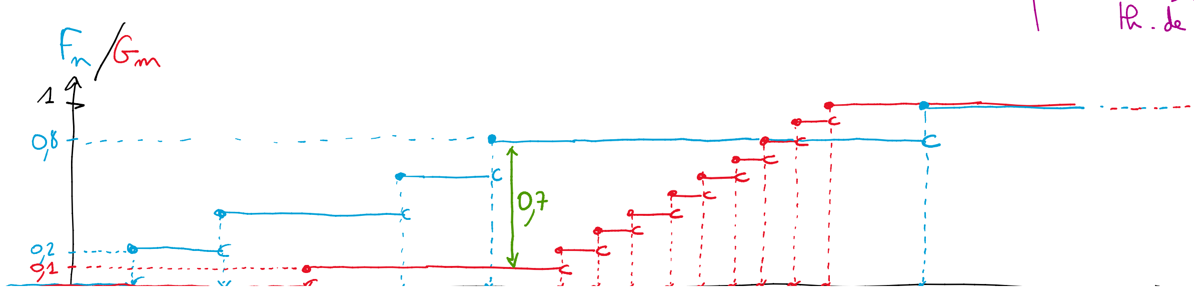
donc on prend une région de rejet de la forme $R = [t, +\infty[$.

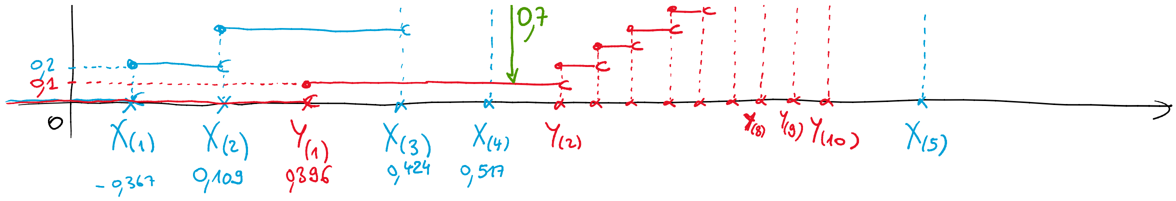
Un choix naturel est $t = d_{n,m,1-\alpha}$ si on veut que le test soit de niveau α.

Sous H_1 par l'inégalité triangulaire,
$$\big| \| F_n - G_m \|_\infty - \| F_X - F_Y \|_\infty \big| \le \underbrace{\| F_n - F_X \|_\infty}_{\to 0} + \underbrace{\| G_m - F_Y \|_\infty}_{\to 0}$$
Th. de Glivenko-Cantelli

$n = 5$, $m = 10$, $\alpha = 5\%$: d'après la table A.16, $d_{5,10,95\%} = 0{,}7$.

Calcul : $h(X,Y)^{obs} =$ ______





donc $\quad h(X,Y)^{obs} = 0{,}7$.

$h(X,Y)^{obs} \in [0{,}7 ; +\infty[$ donc on rejette H0 : les échantillons suivent des lois différentes.

$h(x,y) \in \{0, \frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}, 1\}$.

donc $\quad d_{n,m,\alpha} = \frac{1}{10} \inf \left\{ i \in \{0,\dots,10\} : \overbrace{\mathbb{P}\left(h(X,Y) \geq \frac{i}{10}\right)}^{11 \text{ valeurs à calculer}} \geq \alpha \right\}$