COMPLÉMENT II

Statistiques d'ordre

(*Biblio : Cottrell p. 53*) On considère n variables aléatoires $(X_1, \dots, X_n)$ indépendantes de même loi de densité f sur $\mathbb{R}$, de fonction de répartition F.

Proposition 1. *Les $(X_i)_{1\leq i\leq n}$ sont presque sûrement tous distincts.*

DÉMONSTRATION. En effet soit $i \neq j$. $\mathbb{P}(X_i = X_j) = \int_{\mathbb{R}^2} 1_0(x-y)f(x)f(y)dxdy = \int_{\mathbb{R}} f(x)(\int_{\mathbb{R}} 1_x(y)f(y)dy)dx$ d'après Fubini. La deuxième intégrale est nulle car le singleton $\{x\}$ est de mesure de Lebesgue nulle. On en déduit
$\mathbb{P}(\exists(i \neq j), X_i = X_j) = \mathbb{P}(\cup_{i\neq j}(X_i = X_j)) \leq \sum_{i\neq j} \mathbb{P}(X_i = X_j) = 0$ □

Par conséquent on peut définir pour presque tout $\omega \in \Omega$ la permutation $\sigma(\omega)$ telle que
$X_{\sigma(\omega)(1)}(\omega) < X_{\sigma(\omega)(2)}(\omega) < \cdots < X_{\sigma(\omega)(n)}(\omega)$.

Proposition 2. *σ est une variable aléatoire à valeurs dans l'ensemble des permutations S_n, de loi uniforme sur S_n.*

DÉMONSTRATION. On commence par montrer que σ est bien une variable aléatoire : puisque S_n est un ensemble fini, il suffit de montrer que si τ est une permutation, $\sigma^{-1}(\{\tau\})$ est mesurable. Or
$\sigma^{-1}(\{\tau\}) = \{\omega \in \Omega, X_{\tau(1)}(\omega) < X_{\tau(2)}(\omega) < \dots X_{\tau(n)}(\omega)\} = \{\omega \in \Omega, (X_1, \dots, X_n) \in B_\tau(\omega)\}$, où
$B_\tau = \{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n, x_{\tau(1)} < x_{\tau(2)} < \cdots < x_{\tau(n)}\}$. B_τ est bien un borélien de $\mathbb{R}^n$, par conséquent $\{\omega \in \Omega, (X_1, \dots, X_n) \in B_\tau\}$ est bien mesurable et σ est donc bien une variable aléatoire.
Pour vérifier que σ a bien une loi uniforme, on va montrer que si τ est une permutation quelconque de S_n, $\mathbb{P}(\sigma = \tau)$ ne dépend pas de τ. En effet
$\mathbb{P}(\sigma = \tau) = \mathbb{P}(\{\omega \in \Omega, X_{\tau(1)}(\omega) < X_{\tau(2)}(\omega) < \dots X_{\tau(n)}\omega\})$. Or la densité du n-uplet $(X_1, \dots, X_n)$ est $f(x_1)\dots f(x_n)$: elle est invariante par permutation des coordonnées, et le n-uplet $(X_{\tau(1)}, \dots, X_{\tau(n)})$ a donc même loi que $(X_1, \dots, X_n)$. Par conséquent,
$\mathbb{P}(\{\omega \in \Omega, X_{\tau(1)}(\omega) < X_{\tau(2)}(\omega) < \dots X_{\tau(n)}\omega\}) = \mathbb{P}(\{\omega \in \Omega, X_1(\omega) < X_2(\omega) < \dots X_n(\omega)\})$
ne dépend pas de τ. □

On note souvent pour simplifier les notations $X_{(i)}(\omega) = X_{\sigma(\omega)(i)}$. $X_{(i)}$ est appelé statistique d'ordre de rang i de l'échantillon $(X_1, \dots, X_n)$. En particulier, $X_{(1)} = \min(X_i)$ et $X_{(n)} = \max(X_i)$. Également souvent étudiée, si n est impair, $X_{(n+1)/2}$ est la médiane des X_i. Il est facile de déterminer la loi de $X_{(1)}$ et celle de $X_{(n)}$. Le résultat se généralise :

Proposition 3.
- *Le n-uplet $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ a pour densité $n!1_{x_1<x_2<\cdots<x_n}\prod_{i=1}^n f(x_i)$*
- *Pour tout $1 \leq i \leq n$, $X_{(i)}$ a pour densité $i\binom{n}{i}f(x)(F(x))^{i-1}(1-F(x))^{n-i}$*

DÉMONSTRATION. – En effet, si B est un borélien de $\mathbb{R}^n$,
$\mathbb{P}((X_{(1)}, X_{(2)}, \dots, X_{(n)}) \in B) = \sum_{\tau \in S_n} \mathbb{P}((X_{(1)}, X_{(2)}, \dots, X_{(n)}) \in B \text{ et } \sigma = \tau)$
$= \sum_{\tau \in S_n} \mathbb{P}((X_{\tau(1)}, X_{\tau(2)}, \dots, X_{\tau(n)}) \in B \text{ et } (X_{\tau(1)} < X_{\tau(2)} < X_{\tau(n)}))$
Or si τ est une permutation de S_n, $(X_{\tau(1)} < X_{\tau(2)} < X_{\tau(n)})$ a même loi que $(X_1, \dots, X_n)$. Ainsi chacun des $n!$ événements de la somme a la même probabilité
$\mathbb{P}((X_1, X_2, \dots, X_n) \in B \text{ et } (X_1 < X_2 < \cdots < X_n)) = \int_B 1_{x_1<x_2<\cdots<x_n} f(x_1) \dots f(x_n) dx_1 \dots dx_n$. La densité de $(X_{(1)}, X_{(2)}, \dots, X_{(n)}$ est donc bien $n! 1_{x_1<x_2<\cdots<x_n} f(x_1) \dots f(x_n)$.

– Une fois qu'on a la densité du n-uplet, on pourrait trouver la densité d'une coordonnée en intégrant sur les autres : ainsi si $1 \leq i \leq n$, la densité de $X_{(i)}$ est :
$n! \int_{\mathbb{R}^{n-1}} 1_{x_1<x_2<\cdots<x_n} f(x_1) \dots f(x_n) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_n$
Néanmoins il est plus simple de s'inspirer du calcul classique dans les cas particuliers $i = 1$ ou $i = n$ et de calculer la fonction de répartition de $X_{(i)}$: si $x \in \mathbb{R}$, l'événement $\{X_{(i)} \leq x\}$ est la réunion (disjointe) pour tous les $k \geq i$ des événements "Il y a exactement k éléments du n-uplets inférieurs à x et $n-k$ supérieurs strictement à x". De plus, si $K \subset \{1, \dots, n\}$ vérifie $|K| = k$,
$\mathbb{P}(\forall j \in K, X_j \leq x \text{ et } \forall j \notin K, X_j > x) = F(x)^k (1 - F(x))^{n-k}$ ne dépend pas de K, mais uniquement de son cardinal k. Puisqu'il y a $\binom{n}{k}$ parties de cardinal k, on a finalement

$$\mathbb{P}(X_{(i)} \leq x) = \sum_{k=i}^n \binom{n}{k} F(x)^k (1 - F(x))^{n-k})$$

Pour trouver la densité, il ne reste plus qu'à dériver par rapport à x :
$f_{X_{(i)}}(x) = f(x) \sum_{k=i}^n k \binom{n}{k} F(x)^{k-1} (1 - F(x))^{n-k}) - f(x) \sum_{k=i}^n (n-k) \binom{n}{k} F(x)^k (1 - F(x))^{n-k-1})$
En utilisant les égalités $k\binom{n}{k} = n\binom{n-1}{k-1}$ et $(n-k)\binom{n}{k} = n\binom{n-1}{k}$ il vient
$f_{X_{(i)}}(x) = nf(x) \sum_{k=i}^n \binom{n-1}{k-1} F(x)^{k-1} (1 - F(x))^{n-k}) - nf(x) \sum_{k=i}^n \binom{n-1}{k} F(x)^k (1 - F(x))^{n-k-1})$
Il reste à effectuer un changement de variable dans la première somme $j = k - 1$ et à simplifier pour obtenir enfin :
$f_{X_{(i)}}(x) = nf(x) \binom{n-1}{i-1} F(x)^{i-1} (1 - F(x))^{n-i} = i \binom{n}{i} f(x) F(x)^{i-1} (1 - F(x))^{n-i}.$

□

Deux cas particuliers sont souvent regardés :

(1) Cas où les X_i sont uniformes sur $[0, 1]$. La densité du n-uplet $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ est alors $n! 1_{0<x_1<x_2<\cdots<x_n<1}$: il s'agit d'une loi uniforme sur le simplexe $\{0 < x_1 < x_2 < \cdots < x_n < 1\}$. De plus, pour tout $1 \leq i \leq n$, $X_{(i)}$ a alors pour densité sur $[0, 1]$ $i\binom{n}{i} x^{i-1} (1 - x)^{n-i}$: il s'agit d'une loi β.

(2) Cas où les X_i sont exponentielles de paramètre 1. La densité du n-uplet $(X_{(1)}, X_{(2)}, \dots, X_{(n)})$ est alors $n! 1_{0<x_1<x_2<\cdots<x_n} e^{-\sum_{i=1}^n x_i}$ On peut remarquer que cette densité se transforme simplement avec le difféomorphisme de $\{0 < x_1 < \dots, x_n\}$ dans $(\mathbb{R}^+)^n$ $(x_1, \dots, x_n) \mapsto (x_1, x_2 - x_1, \dots, x_n - x_{n-1})$. Ce difféomorphisme est de jacobien 1, et de réciproque
$(y_1, y_2, \dots y_n) \mapsto (y_1, y_1 + y_2, \dots, \sum_{i=1}^n y_i)$. On en déduit finalement que le n-uplet $(Y_1, \dots, Y_n)$ défini par $Y_i = X_{(i+1)} - X_{(i)}$ a pour densité $n! \prod_{i=1}^n 1_{\mathbb{R}^+}(y_i) \prod_{i=1}^n e^{-(n-i+1)y_i}$: en particulier les Y_i sont indépendantes, de loi exponentielle de paramètre respectif $n - i + 1$.

De plus, pour tout $1 \leq i \leq n$, $X_{(i)}$ a alors pour densité sur $\mathbb{R}^+$ $i\binom{n}{i} e^{-(n+1-i)x} (1 - e^{-x})^{i-1}$. On retrouve en particulier que le minimum $X_{(1)}$ suit une loi exponentielle de paramètre n.