DM 1

Exercice 1 *Réordonnement*

Soit $n \in \mathbb{N}$, on pose

$$g_n : \begin{cases} \mathbb{R}^n & \to \mathbb{R}^n \\ (x_1, \ldots, x_n) & \mapsto (x_{\sigma(1)}, \ldots, x_{\sigma(n)}) \text{ où } \sigma \in \mathfrak{S}_n \text{ est telle que } x_{\sigma(1)} \leq x_{\sigma(2)} \leq \ldots \leq x_{\sigma(n)} \end{cases}.$$

la fonction de réordonnement des coordonnées des vecteurs de $\mathbb{R}^n$ par ordre croissant. Soit maintenant $(X_i)_{i\geq 1}$ une suite de variables aléatoires réelles i.i.d. Pour tout $n \in \mathbb{N}$, on pose

$$(Y_1^{(n)}, \ldots, Y_n^{(n)}) := g_n(X_1, \ldots, X_n)$$

le vecteur des n premières valeurs de la suite $(X_i)_{i\geq 0}$ ré-ordonnées.

1. (a) Montrer que pour tout $k \in [\![1; n]\!]$ et tout $x \in \mathbb{R}$, on a que p.s.

$$Y_k^{(n)} \leq x \iff \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}} \geq k.$$

   (b) On suppose que F_X la fonction de répartition de X est strictement croissante sur $\mathbb{R}$, et l'on note pour tout $t \in [0; 1]$,

$$G_X(t) = \inf\{x : F(x) \geq t\}$$

   la fonction quantile de X (avec la convention $\inf \emptyset = +\infty$, et en convenant que la borne inférieure d'une partie non minorée de $\mathbb{R}$ est $-\infty$). Montrer que pour tout $t \in [0; 1]$,

$$Y_{\lfloor tn \rfloor}^{(n)} \underset{n\to\infty}{\longrightarrow} G(t) \quad \text{p.s. ,}$$

   où $\lfloor x \rfloor$ désigne la partie entière de x.

2. On suppose désormais que $X_i \sim \mathcal{U}([0; 1])$, et l'on fixe $n \in \mathbb{N}^*$.
   (a) Montrer que presque sûrement, les $(Y_k(n))_{1\leq k\leq n}$ sont deux à deux distincts.
   (b) Soit $\sigma \in \mathfrak{S}_n$, calculer $\mathbb{P}((Y_1^{(n)}, \ldots, Y_n^{(n)}) = (X_{\sigma(1)}, \ldots, X_{\sigma(n)}))$.
   (c) Montrer que pour toute fonction borélienne bornée $h : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbb{E}[h(Y_1^{(n)}, \ldots, Y_n^{(n)})] = n! \int_{\mathbb{R}^n} h(y_1, \ldots, y_n) \mathbf{1}_{\{0\leq y_1 < \cdots < y_n \leq 1\}} \, dy_1 \ldots dy_n$$

   (d) En déduire que pour tout $k \in [\![1; n]\!]$, la variable aléatoire $Y_k^{(n)}$ admet une densité, que l'on calculera explicitement.

t.s.v.p.

Exercice 2 *Estimation*

Soit $(X_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d., de loi inconnue ν, supposée intégrable. On pose $m := \mathbb{E}[X_1]$ l'espérance de cette loi.

Un estimateur de m est une variable aléatoire μ_n pour un certain $n \geq 1$ telle que

(i) La variable aléatoire μ_n s'écrit comme une fonction des n premières observations : il existe $f_n : \mathbb{R}^n \to \mathbb{R}$ telle que $\mu_n = f_n(X_1, \dots, X_n)$, et f_n ne dépend pas de ν (en particulier on ne peut pas prendre $f_n \equiv m$ et donc $\mu_n = m$, ce qui serait trop facile). Cette condition est nécessaire pour pouvoir utiliser l'estimateur en pratique.

(ii) La probabilité $\mathbb{P}(|\mu_n - m| > x)$ peut être dominée indépendamment de ν.

Soit maintenant $\alpha \in]0, 1[$. Supposons avoir trouvé un nombre déterministe $\delta(\alpha, n)$, indépendant de ν, tel que

$$\mathbb{P}(|\mu_n - m| > \delta(\alpha, n)) \leq \alpha.$$

On dit alors que $I_{\alpha,n} = [\mu_n - \delta(\alpha, n), \mu_n + \delta(\alpha, n)]$ est un intervalle de confiance pour m au seuil $1 - \alpha$. En effet on a

$$\mathbb{P}(m \in I_{n,\alpha}) \geq 1 - \alpha$$

Remarquons que $I_{\alpha,n}$ est un intervalle aléatoire (il dépend de $X_1, \dots X_n$ à travers μ_n). L'évènement $\{m \in I_{n,\alpha}\}$ est un peu étrange (un nombre déterministe appartient à un intervalle aléatoire), mais tout-à-fait bien défini.

1. On pose $\mu_n = \frac{1}{n}(X_1 + \dots X_n)$. Supposons d'abord que ν a variance finie, bornée par un réel supposé connu σ^2. Déduire de l'inégalité de Bienaymé-Tchébychev un intervalle de confiance au niveau α.
2. Supposons maintenant que notre loi inconnue est bornée : il existe $a, b \in \mathbb{R}$ tels que presque sûrement $X_1 \in [a, b]$. On admettra l'inégalité déterministe suivante : pour $\theta \in [0, 1]$ et $u > 0$, on a $\theta e^{-(1-\theta)u} + (1 - \theta)e^{\theta u} \leq e^{u^2/8}$.
   (a) Utiliser cette inégalité pour borner $\mathbb{E}[e^{\lambda(X_1 - m)}]$ et en déduire l'inégalité suivante (où $(\cdot)_+$ désigne la partie positive)
   $$\mathbb{P}(\mu_n - m \geq x) \leq \exp(-2n(x)_+^2/(b - a)^2).$$
   (b) En déduire un intervalle de confiance au niveau α.
3. Supposons que ν est une loi de Bernoulli de paramètre inconnu : comparer les intervalles de confiance obtenus aux questions 1 et 2, d'abord pour $\alpha = 0.05$, puis pour $\alpha \to 0$.
4. On va maintenant construire un estimateur sous la seule hypothèse que la variance de ν est bornée par σ^2, mais dont la largeur quand $\alpha \to 0$ est du même ordre que celle obtenue à la question 2. Soit k a déterminer plus tard et $\ell = \lfloor n/k \rfloor$. Pour $i \geq 1$, on pose
   $$\mu^{(i)} = \frac{1}{\ell} \sum_{j=\ell(i-1)+1}^{\ell i} X_j.$$
   On pose ensuite $\nu_n = \mathrm{Med}(\mu^{(1)}, \dots, \mu^{(k)})$, c'est-à-dire que
   $$\nu_n := \min\{M \in \mathbb{R} : \#\{i, 1 \leq i \leq k : \mu^{(i)} \leq M\} \geq \#\{i, 1 \leq i \leq k : \mu^{(i)} > M\}\}$$
   désigne la plus petite des médianes de l'ensemble $(\mu^{(1)}, \dots, \mu^{(k)})$.
   (a) Montrer que $\#\{i, 1 \leq i \leq k : \mu^{(i)} \leq \nu_n\} \geq k/2$ et $\#\{i, 1 \leq i \leq k : \mu^{(i)} \geq \nu_n\} \geq k/2$.
   (b) Montrer que si $|\nu_n - m| > x$, alors
   $$\#\{i, 1 \leq i \leq k \text{ t.q. } |\mu^{(i)} - m| > x\} \geq k/2.$$

(c) Combiner Bienaymé-Tchébychev et l'inégalité démontrée en question 2 pour donner une estimation de $\mathbb{P}(|\nu_n - m| > x)$.

(d) On suppose $\alpha \leq 1/e^2$ et on s'autorise à choisir k dépendant de α (c'est à dire que le mode de calcul de l'estimateur change selon la précision visée). Trouver un choix de k tel que pour une constante C universelle et $n \geq k$,

$$\mathbb{P}\left(|\nu_n - m| > \sigma\sqrt{\frac{C\log(1/\alpha)}{n}}\right) \leq \alpha.$$

5. Un vieux stastisticien rabat-joie s'exclame : « à quoi sert tout ce charabia ? dans la vraie vie on prend $\alpha = 0,05$ et n ne dépasse jamais 1000 ! » Que lui répondre ?