TD 06 – Inégalité de Chernoff (Suite)

Exercice 1. *Sondage*

Nous sommes en période de campagne BDE à l'ENS de Lyon et nous voulons faire un sondage d'opinion pour estimer la proportion p de la population normalienne souhaitant voter pour la Zicliste. Supposons que l'on interroge n personnes choisies uniformément et indépendamment au hasard, et que chacune d'elle réponde par "Oui, je souhaite voter pour eux" ou "Non, je ne suis souhaite pas voter pour eux". Étant donnés $\theta > 0$ et $0 < \delta < 1$, on souhaite trouver une estimation $\overline{X}$ de p telle que

$$\mathbf{P}\left\{|\overline{X} - p| \leq \theta\right\} > 1 - \delta \, .$$

Par exemple pour $1 - \delta = 0.95$, on pourra ainsi dire que le sondage a une précision de θ à 95%.

1. Que choisir comme estimation $\overline{X}$ de p ?
2. Combien de personnes doit-on interroger pour que l'estimation $\overline{X}$ vérifie nos conditions ? Autrement dit, donner une borne inférieure sur n en termes de θ et δ. On remarquera que cette borne ne dépend pas de la taille de la population totale.
3. Calculer la valeur de n obtenue grâce à votre borne pour les paramètres $\theta = 0.2$ et $1 - \delta = 95\%$.

Exercice 2. *blackjack*

Vous êtes le croupier dans une partie de blackjack au Gala de l'ENS de Lyon, et vous soupçonnez un joueur de tricher en comptant les cartes. En effet, sur les quelques premières mains que vous venez de le voir jouer, il gagne 55% du temps (alors, que, sans tricher, la probabilité de gagner un main est $1/2$). Cependant, vous voulez attendre d'avoir un peu plus de certitude avant de démasquer le joueur.

1. On suppose que le joueur continue de gagner 55% du temps. Combien de mains devez-vous le laisser jouer avant d'être sûr à 90% qu'il triche ?

Exercice 3. *Sous-gaussiennes*

Une variable aléatoire X est dite *sous-gaussienne* de paramètre σ si elle vérifie l'inégalité suivante :

$$\forall \lambda \in \mathbb{R}, \mathbf{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 \sigma^2}{2}} \, .$$

Soit X_1 et X_2 deux variables aléatoires indépendantes, respectivement sous-gaussiennes de paramètre σ_1 et σ_2.

1. Montrez que $X_1 + X_2$ est sous-gaussienne de paramètre $\sqrt{\sigma_1^2 + \sigma_2^2}$. Montrez que cX_1 est $|c|\sigma_1$-sous-gaussienne pour tout $c \in \mathbb{R}$.
2. Montrez que si X est une variable aléatoire σ-sous-gaussienne, alors pour tout $t \geq 0$,

$$\mathbf{P}\left\{X \geq t\right\} \leq e^{-\frac{t^2}{2\sigma^2}}$$

.

3. Soit μ un réel et $X_1, \ldots, X_n$ des variables aléatoires telles que les $X_i - \mu$ sont indépendantes et σ-sous-gaussiennes. Soit $\delta \in [0, 1]$. Montrez qu'avec probabilité au moins $1 - \delta$, on a

$$\mu \leq \hat{\mu} + \sqrt{\frac{\sigma^2 \log(1/\delta)}{2n}} \, ,$$

où $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Exercice 4. *Algorithme probabiliste pour calculer la médiane*

On étudie un algorithme probabiliste [1] pour déterminer la médiane d'un ensemble $E = \{x_1, \dots, x_n\}$ de n nombres réels en temps $O(n)$. On rappelle que m est une médiane de E si au moins $\lceil n/2 \rceil$ des élements de E sont inférieurs ou égaux à m, et au moins $\lceil n/2 \rceil$ des élements de E sont supérieurs ou égaux à m. Pour simplifier on suppose n impair (ce qui fait que la médiane est unique) et on suppose aussi que les éléments de E sont tous distincts.

Voici comment fonctionne l'algorithme

(a) Soit $(Y_i)_{1 \le i \le n}$ une suite de v.a. i.i.d. de loi de Bernoulli de paramètre $n^{-1/4}$. On considère le sous-ensemble aléatoire de E défini par $F = \{x_i \,:\, Y_i = 1\}$. Si card $F \le \frac{2}{3} n^{3/4}$ ou card $F \ge 2n^{3/4}$ on répond «`ERREUR 1`».

(b) On trie F et on appelle d le $\lfloor \frac{1}{2} n^{3/4} - \sqrt{n} \rfloor$ème plus petit élément de F, et u le $\lfloor \frac{1}{2} n^{3/4} - \sqrt{n} \rfloor$ème plus grand élément de F.

(c) On détermine le rang de d et de u dans E (l'élément minimal a rang 1, l'élément maximal a rang n), que l'on note respectivement r_d et r_u. Si $r_d > n/2$ ou $r_u < n/2$ on répond «`ERREUR 2`».

(d) On note $G = \{x_i \in E \,:\, d < x_i < u\}$. Si card $G \ge 4n^{3/4}$ on répond «`ERREUR 3`».

(e) On trie G et on renvoie le $(\lceil n/2 \rceil - r_d)$ème élement de G.

1. Justifier pourquoi l'algorithme retourne la médiane en temps $O(n)$ lorsqu'il ne répond pas de message d'erreur.

2. Montrer que pour $i \in \{1, 2, 3\}$, on a

$$\lim_{n \to \infty} \Pr\left(\text{l'algorithme retourne «}\texttt{ERREUR } i\text{»}\right) = 0.$$

Pour simplifier l'analyse et éviter d'écrire des symobles $\lfloor \cdot \rfloor$ ou $\lceil \cdot \rceil$, on pourra supposer implicitement que des nombres tels que $\sqrt{n}$, $\frac{1}{2} n^{3/4}$, $\dots$ sont des entiers

Exercice 5. *Graphe Aléatoire Bipartite*

Soit $0 < p < 1$ et $n \in \mathbb{N}^*$. On définit un graphe aléatoire non orienté $H_{2n,p}$ de la manière suivante. On se donne une famille $\{X_{i,j} \,:\, 1 \le i \le n,\ n+1 \le j \le 2n\}$ de v.a. i.i.d. de loi de Bernoulli de paramètre p. On pose alors $H_{2n,p} = (V, E)$, avec $V = \{1, \dots, 2n\}$ et

$$E = \{(i, j) \,:\, X_{i,j} = 1\} \subset \{1, \dots, n\} \times \{n+1, \dots, 2n\}.$$

1. Quelle est la loi du nombre d'arêtes de $H_{2n,p}$?

2. Quelle est l'espérance du nombre de sommets isolés de $H_{2n,p}$?

3. Dans cette question on pose $p = c \log(n)/n$ pour un nombre réel $c > 0$.

1. Montrer que si $c > 1$, alors

$$\lim_{n \to \infty} \mathsf{P}(H_{2n,p} \text{ a un sommet isolé}) = 0.$$

2. Montrer que si $c < 1$, alors

$$\lim_{n \to \infty} \mathsf{P}(H_{2n,p} \text{ a un sommet isolé}) = 1.$$

4. Dans cette question on pose $p = 1/2$. Montrer qu'il existe une constante $C > 0$ telle que

$$\lim_{n \to \infty} \mathsf{P}\left(\text{tous les sommets de } H_{2n,p} \text{ ont un degré inférieur à } \frac{n}{2} + C\sqrt{n \log n}\right) = 1.$$

1. Remarque : il existe un algorithme déterministe de même performance