Devoir

Exercice 1 : problème du collectionneur (4 pts)

Dans chaque paquet de céréales, une vignette est offerte. Il y en a trois différentes à collectionner.

1. Après avoir obtenu la première vignette, combien de paquets en moyenne doit-on acheter pour obtenir une vignette différente ?
   Indication : on reconnaîtra une loi usuelle.
2. Combien ensuite faudra-t-il acheter de paquets en moyenne pour obtenir la dernière vignette ?
3. Généraliser au cas de n vignettes : quel est le nombre moyen de paquets qu'il faut acheter pour avoir une collection complète de n vignettes ?

Exercice 2 : méthode des moindres carrés (16 pts)

On sait que deux grandeurs x et y sont proportionnelles : $y = ax$ avec $a \in \mathbf{R}$. On cherche à déterminer la valeur de a.

On s'est donné deux valeurs pour x : $x_1 = 2$ et $x_2 = 4$. Puis on a mesuré les valeurs y_1 et y_2 correspondantes. Mais ces mesures sont perturbées par un bruit supposé aléatoire. Leurs valeurs sont ainsi décrites par des variables aléatoires

$$Y_1 = 2a + \mathcal{E}_1 \quad \text{et} \quad Y_2 = 4a + \mathcal{E}_2,$$

où $\mathcal{E}_1$ et $\mathcal{E}_2$ sont deux variables indépendantes de même densité f définie sur $\mathbf{R}$, d'espérance nulle et de variance 1 : $\mathbf{E}(\mathcal{E}_i) = 0$ et $\mathbf{V}(\mathcal{E}_i) = 1$.

I. **Estimation de a** (5 pts)
   On cherche à estimer a à partir des valeurs obtenues pour Y_1 et Y_2. Comme a représente le rapport entre y et x, on pourrait logiquement estimer a par les rapports $\frac{Y_1}{x_1}$ et $\frac{Y_2}{x_2}$, ou pourquoi pas, par la moyenne des deux : $\frac{1}{2}(\frac{Y_1}{2} + \frac{Y_2}{4})$. Mais est-ce la meilleure façon de procéder ?

   On propose d'estimer a par la valeur

$$A = \alpha Y_1 + \beta Y_2,$$

   où α et β sont des nombres réels. On impose les conditions suivantes pour que l'estimation soit la meilleure possible :

$$\mathbf{E}(A) = a \quad \text{et} \quad \mathbf{V}(A) \text{ est minimal.}$$

1. Interpréter ces conditions. Pourquoi sont-elles naturelles pour espérer obtenir une bonne estimation de a ?
2. Montrer que les conditions impliquent que $\alpha = \frac{1}{10}$ et $\beta = \frac{1}{5}$.
3. On obtient $\beta > \alpha$. Cela signifie que pour estimer a, on donne plus de poids à la mesure Y_2 qu'à la mesure Y_1. Essayer d'expliquer pourquoi c'est judicieux pour améliorer l'estimation de a.

Remarque : les valeurs obtenues pour α et β correspondent exactement aux valeurs qu'on obtient quand on utilise la méthode des moindres carrés pour obtenir la meilleure droite linéaire à partir des deux points $(2, Y_1)$ et $(4, Y_2)$ obtenus. Notre étude justifie que cette méthode est en un sens la meilleure.

II. **Loi de A** (5 pts)
On considère désormais la variable $A = \frac{1}{10}Y_1 + \frac{1}{5}Y_2$ obtenue dans la partie précédente.

1. Supposons pour commencer que $\mathcal{E}_1$ et $\mathcal{E}_2$ suivent la loi normale $\mathcal{N}(0, 1)$. Quelle est alors la loi de A ?
2. Déterminer dans ce cas la valeur de η pour laquelle $\mathbf{P}(|A - a| \leqslant \eta) \approx 0,99$.
3. Dans le cas général, on détermine la loi de A en utilisant la densité f de $\mathcal{E}_1$ et $\mathcal{E}_2$. Soit $t \in \mathbf{R}$. Déterminer l'expression de la fonction de répartition $F_A(t)$ sous forme intégrale en fonction de la densité f.

III. **Généralisation** (6 pts)
On effectue maintenant n mesures en n points $x_1, \ldots, x_n$. On obtient à chaque fois une valeur $Y_i = ax_i + \mathcal{E}_i$ où la variable $\mathcal{E}_i$ a les mêmes caractéristiques que $\mathcal{E}_1$. La méthode des moindre carrés permet d'estimer a par la variable

$$A = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

On aimerait déterminer le nombre de points n qu'il faut considérer pour assurer avec un risque de 1% que $|A - a| < 10^{-2}$.

1. Donner $\mathbf{E}(A)$ et $\mathbf{V}(A)$.
2. Pour quelle raison les variables $x_i Y_i$ ne satisfont pas les hypothèses du théorème central limite ?
3. À l'aide de l'inégalité de Bienaymé-Tchebychev, donner une condition sur les x_i qui garantisse la précision demandée.

   Afin de pouvoir utiliser le TCL, nous allons simplifier notre problème en choisissant des nombres x_i tous égaux à x. La variable A devient alors $A = a + \frac{1}{nx} \sum_{i=1}^n \mathcal{E}_i$.
4. À l'aide du théorème central limite, déterminer une condition sur n et x qui permette d'assurer la précision demandée.
   Qu'obtient-on pour $x = 2$ et $x = 4$?