DS 9 (B)

Devoir surveillé du Jeudi 27 Février

La calculatrice est interdite. Durée : 4h

Est-il possible que le marketing digital pose des problèmes de sécurité des données personnelles ? De récents travaux[1], mettant en cause les outils de mesure de performance en temps réel des différentes campagnes de publicité sur internet, démontrent que certaines données très sensibles (préférences religieuses, sexuelles, etc.) peuvent être obtenues par des segmentations précises des audiences et sans aucune action de la part de l'utilisateur.
Dans ce problème, nous nous intéressons à une méthode proposée pour protéger ces données, méthode baptisée **confidentialité différentielle**.

Les parties I et II sont totalement indépendantes. Vous trouverez une aide Python en fin de sujet.

On considère un espace probabilisé $(\Omega, \mathcal{A}, P)$ sur lequel sont définies les variables aléatoires qui apparaissent dans l'énoncé.

Partie I - Lois de Laplace - propriétés et simulation

Soit $\alpha \in \mathbb{R}$ et $\beta > 0$. On dit qu'une variable aléatoire réelle à densité suit une loi de Laplace de paramètre (α, β), notée $\mathcal{L}(\alpha, \beta)$, si elle admet comme densité la fonction f donnée par :

$$\forall t \in \mathbb{R}, \quad f(t) = \frac{1}{2\beta} \exp\left(-\frac{|t-\alpha|}{\beta}\right).$$

1. Vérifier que f est bien une densité de probabilité d'une variable aléatoire réelle.
2. Déterminer la fonction de répartition, notée Ψ, de la loi $\mathcal{L}(0,1)$.
3. On suppose que X suit la loi $\mathcal{L}(0,1)$.
   (a) Montrer que $\beta X + \alpha$ suit la loi $\mathcal{L}(\alpha, \beta)$.
   (b) En déduire la fonction de répartition de la loi $\mathcal{L}(\alpha, \beta)$.
4. *Espérance et variance.*
   (a) On suppose que X suit la loi $\mathcal{L}(0,1)$.
   Montrer que $E(X)$ et $V(X)$ existent et valent respectivement 0 et 2.
   (b) En déduire l'existence et les valeurs de l'espérance et de la variance d'une variable aléatoire réelle qui suit la loi $\mathcal{L}(\alpha, \beta)$.
5. *Simulation à partir d'une loi exponentielle.*
   Soit U une variable aléatoire qui suit la loi exponentielle de paramètre 1 et V une variable aléatoire qui suit la loi de Bernoulli de paramètre $\frac{1}{2}$ et indépendante de U.
   (a) En utilisant le système complet naturellement associé à V, montrer que $X = (2V-1)U$ suit la loi $\mathcal{L}(0,1)$.
   (b) Compléter la définition `Python` ci-dessous pour que la fonction ainsi définie réalise la simulation d'une variable aléatoire qui suit la loi $\mathcal{L}(\alpha, \beta)$:

[1] Par exemple, A. Korolova. Privacy violations using microtargeted ads: A case study (2010)

```
def Laplace(alpha,beta):
  if ..... <= 1/2:
    V = 1
  else:
    V = 0
  X = (2*V - 1) * rd.exponential(1)
  return( ..... )
```

Partie II - Lois ε-différentielles

Soit $\varepsilon > 0$. On dit que (X, Y), un couple de variables aléatoires, est un couple ε-différentiel si, pour tout intervalle I de $\mathbb{R}$:

$$e^{-\varepsilon} P([X \in I]) \leq P([Y \in I]) \leq e^{\varepsilon} P([X \in I]).$$

Intuitivement, les lois de X et Y seront d'autant plus proches que le plus petit ε tel que (X, Y) soit un couple ε-différentiel est proche de 0.

6. Soit (X, Y, Z) un triplet de variables aléatoires réelles.
   (a) Montrer que si (X, Y) est ε-différentiel alors (Y, X) l'est aussi.
   (b) Montrer que si (X, Y) est ε-différentiel et (Y, Z) est ε'-différentiel alors (X, Z) est $(\varepsilon + \varepsilon')$-différentiel.

7. Soit (X, Y) un couple de variables aléatoires réelles discrètes. On suppose que $X(\Omega) \cup Y(\Omega) = \{z_n \mid n \in J\}$ où J est un sous ensemble non vide de $\mathbb{N}$.
   Montrer que (X, Y) est ε-différentiel si et seulement si

$$\forall n \in J, \quad e^{-\varepsilon} P([X = z_n]) \leq P([Y = z_n]) \leq e^{\varepsilon} P([X = z_n]).$$

8. *Premier exemple.*
   Dans cette question, on suppose que X suit la loi géométrique de paramètre $\frac{1}{2}$, Z suit la loi de Bernoulli de paramètre $p \in]0, 1[$ et elles sont indépendantes. On pose $Y = X + Z$.
   (a) Déterminer la loi de Y.
   (b) Établir que pour tout $k \in \mathbb{N}^*$, $1 - p \leq \dfrac{P([Y = k])}{P([X = k])} \leq \dfrac{1}{1 - p}$.
   (c) En déduire que (X, Y) est $-\ln(1 - p)$-différentiel.
   (d) Que se passe-t-il lorsque p s'approche de 0 ou lorsqu'il s'approche de 1 ? Était-ce prévisible ?

9. On suppose que X et Y sont deux variables à densité de densités respectives f et g et de fonction de répartition F et G.
   (a) On suppose que pour tout $t \in \mathbb{R}$, $e^{-\varepsilon} f(t) \leq g(t) \leq e^{\varepsilon} f(t)$.
   Montrer que (X, Y) est ε-différentiel.
   (b) On suppose dans la suite de cette question que (X, Y) est ε-différentiel.
   Soit $h > 0$ et $t \in \mathbb{R}$ où f et g sont continues.

   Montrer que :

$$e^{-\varepsilon} \frac{F(t + h) - F(t)}{h} \leq \frac{G(t + h) - G(t)}{h} \leq e^{\varepsilon} \frac{F(t + h) - F(t)}{h}$$

   En conclure que : $e^{-\varepsilon} f(t) \leq g(t) \leq e^{\varepsilon} f(t)$.

10. *Deuxième exemple : lois de Cauchy.*

    (a) Montrer que $\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t^2+1}\,dt$ converge. On admet que cette intégrale est égale à π.

    (b) On définit, pour $a>0$, la fonction f_a sur $\mathbb{R}$ par, pour tout $t\in\mathbb{R}$, $f_a(t)=\dfrac{a}{\pi(t^2+a^2)}$.
    Montrer que f_a est une densité de probabilité d'une variable aléatoire à densité.

    (c) On suppose que X et Y sont deux variables aléatoires admettant comme densités respectives f_1 et f_a avec $a>1$.
    Montrer que (X,Y) est $\ln(a)$-différentiel.

11. *Une première interprétation.*

    On suppose que (X,Y) est un couple ε-différentiel et que U est une variable de Bernoulli de paramètre $p\in]0,1[$ indépendante de X et Y.

    On définit la variable aléatoire Z par :

$$\forall\omega\in\Omega,\quad Z(\omega)=\begin{cases}X(\omega) & \text{si } U(\omega)=1,\\ Y(\omega) & \text{sinon.}\end{cases}$$

    (a) Soit I un intervalle de $\mathbb{R}$ telle que $P([Z\in I])\neq 0$.
    Montrer que : $P_{[Z\in I]}([U=1])=p\dfrac{P([X\in I])}{pP([X\in I])+(1-p)P([Y\in I])}$.
    En déduire que :

$$\frac{p}{p+(1-p)e^{\varepsilon}}\leq P_{[Z\in I]}([U=1])\leq\frac{p}{p+(1-p)e^{-\varepsilon}}.$$

    (b) Si ε est proche de zéro, le fait de disposer d'une information sur la valeur de Z change-t-il notablement le paramètre de la loi de U et par conséquent la probabilité d'en déduire la valeur prise par U ?

Partie III - Confidentialité différentielle

- Soit $d\in\mathbb{N}^*$. On considère $D=[\![0,d]\!]$ et n un entier naturel plus grand que 2.
- On dira que deux éléments de D^n, a et b, sont voisins si ils ne différent que d'une composante au plus. On note $\mathcal{V}$ l'ensemble des couples de voisins.
- On considère q une application de D^n dans $\mathbb{R}$.

Concrètement, un élément de D^n représente une table d'une base de donnée et q une requête sur cette base.

Étant donné $a=(a_1,...,a_n)$, on s'intéresse au problème de la confidentialité de certains des a_i lorsque les autres a_i sont connus, ainsi que D, q et $q(a)$.

12. Dans cette question on suppose que $a_2,...,a_n$ sont connus et on cherche à protéger a_1.

    (a) Quelle est probabilité d'obtenir la bonne valeur de a_1 si l'on choisit une valeur au hasard dans $[\![0,d]\!]$?

    (b) Dans cette question $q(a_1,...,a_n)=\displaystyle\sum_{i=1}^{n}a_i$.
    Montrer que si $q(a)$ est publique alors on sait déterminer la valeur de a_1.

On dit que l'on dispose d'un procédé de ε-confidentialité de D^n pour q si :

(c1) pour tout $a \in D^n$, on dispose d'une variable aléatoire réelle X_a ;

(c2) pour tout $(a,b) \in \mathcal{V}$, (X_a, X_b) est ε-différentiel ;

(c3) pour tout $a \in D^n$, $E(X_a) = q(a)$.

13. *Majoration de la probabilité de trouver a_1.*

Dans cette question, nous allons justifier en partie la terminologie. On suppose à nouveau que $a_2, ..., a_n$ sont connus, que l'on cherche à protéger a_1 et que :

- Le public connaît des d'intervalles $I_0, ..., I_d$ disjoints de réunion $\mathbb{R}$ tels qu'avec les valeurs fixées de $a_2, ..., a_n$, si $q(a) \in I_j$ alors $a_1 = j$. Cela signifie que si $q(a)$ est publique alors a_1 aussi.
- On dispose d'un procédé de ε-confidentialité de D^n pour q et que l'on rend X_a publique à la place de $q(a)$.

On considère alors que l'expérience aléatoire modélisée par $(\Omega, \mathcal{A}, P)$ comporte comme première étape le choix au hasard de a_1 dans $[\![0, d]\!]$ et on définit :

- A_1 la variable aléatoire associée à ce choix.
- pour tout $j \in [\![0, d]\!]$, $Y_j = X_{(j, a_2, ..., a_n)}$. On suppose que A_1 et Y_j sont indépendantes pour tout $j \in [\![0, d]\!]$.
- la variable aléatoire réelle R par : $\forall \omega \in \Omega$, si $A_1(\omega) = j$ alors on détermine l'unique k tel que $Y_j(\omega) \in I_k$ et on pose $R(\omega) = k$.
- $\theta = P([R = A_1])$.

(a) Montrer que : $\theta = \displaystyle\sum_{j=0}^{d} P([Y_j \in I_j] \cap [A_1 = j])$.

(b) En déduire que : $\theta = \dfrac{1}{d+1} \displaystyle\sum_{j=0}^{d} P([Y_j \in I_j])$.

(c) En conclure que :
$$\theta \leq \frac{1}{d+1} \left(e^{\varepsilon} - (e^{\varepsilon} - 1) P([Y_0 \in I_0])\right) \leq \frac{e^{\varepsilon}}{d+1}$$

(d) On pose $\rho = \dfrac{1}{d+1}$ et $\tau = \dfrac{\theta - \rho}{\rho}$.
Donner une majoration de τ. Que représente cette quantité ?
Qu'en déduire concernant la méthode de confidentialité présentée dans cette question lorsque ε est proche de 0 ?

On pose $\delta = \max\limits_{(a,b) \in \mathcal{V}} |q(a) - q(b)|$ et on suppose que $\delta > 0$.

14. Dans cette question, pour tout $a \in D^n$, on pose $X_a = q(a) + Y$ où Y suit la loi de Laplace de paramètre $(0, \beta)$.

(a) Pour tout $a \in D^n$, déterminer $E(X_a)$ et une densité de probabilité f_a de la loi de X_a en fonction de $q(a)$ et de β.

(b) Montrer que pour tout $t \in \mathbb{R}$ et $(a,b) \in \mathcal{V}$, $f_a(t) \leq \exp\left(\dfrac{\delta}{\beta}\right) f_b(t)$.

En déduire que pour tout $(a,b) \in \mathcal{V}$, (X_a, X_b) est $\dfrac{\delta}{\beta}$-différentiel.

(c) Comment choisir β pour disposer alors d'un procédé de ε-confidentialité de D^n pour q ?

15. Dans cette question, pour tout $a = (a_1, ..., a_n)$ appartenant à D^n, $q(a) = \sum_{k=1}^{n} a_k$.

(a) Quelle est la valeur de δ ?

On utilise dans la suite le procédé de ε-confidentialité tel qu'il a été défini dans la question 14 mais au lieu de publier la valeur X_a, on procède ainsi :

- si $X_a < \frac{1}{2}$, on publie 0 ;
- si $X_a \in [k - \frac{1}{2}, k + \frac{1}{2}[$ où $k \in [\![1, nd - 1]\!]$, on publie k ;
- sinon, on publie nd.

(b) Montrer que la valeur aléatoire Z_a publiée vérifie :

$$Z_a = \begin{cases} 0 & \text{si } X_a < \frac{1}{2}, \\ \lfloor X_a + \frac{1}{2} \rfloor & \text{si } X_a \in [\frac{1}{2}, nd - \frac{1}{2}[, \\ nd & \text{si } X_a \geq nd - \frac{1}{2}. \end{cases}$$

(c) Écrire un script qui pour d, n et ε saisis par l'utilisateur, génère une valeur aléatoire de $a \in D^n$ puis affiche $q(a)$ et Z_a.

(d) Pour $n = 1000$, $d = 4$ et ε choisi par l'utilisateur, écrire un script qui estime la valeur moyenne de $\dfrac{|Z_a - q(a)|}{q(a)}$ (on considèrera que $q(a)$ est toujours non nul).

N.B. À titre d'information, on obtient le tableau de valeurs suivant :

ε	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2
Moyenne	1.91 %	1%	0.6 %	0.5 %	0.3 %	0.3 %	0.28 %	0.2 %	0.2 %	0.19 %	0.17 %	0.16 %

Aide Python. La librairie `numpy.random` (avec le raccourci `rd`) permet de simuler, en particulier, les lois exponentielles et uniformes discrètes. Par exemple :

- `rd.exponential(0.5)` renvoie un réel aléatoire qui suit la loi exponentielle d'espérance $0,5$.
- `rd.randint(-1, 4, 10)` renvoie un vecteur aléatoire de taille 10 dont les coefficients sont des variables indépendantes qui suivent la loi uniforme discrètes sur $[\![-1, 3]\!]$.