MATHÉMATIQUES I - ESSEC E 2017

Proposition de corrigé par David Meneu

Lycée Champollion - Grenoble, pour Major-Prépa

Est-il possible que le marketing digital pose des problèmes de sécurité des données personnelles ? De récents travaux, 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** étaient totalement indépendantes.

On considère un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ 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 paramètres (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. La fonction f est tout d'abord bien continue sur tout $\mathbb{R}$ comme composée de telles fonctions, et (strictement) positive sur tout $\mathbb{R}$ par stricte positivité de l'exponentielle et puisque $\beta > 0$.
   Par ailleurs, sous réserve de convergence absolue :

$$\begin{aligned}\int_{-\infty}^{+\infty} f(t)\mathrm{d}t &= \frac{1}{2\beta}\int_{-\infty}^{\alpha} \exp\left(\frac{t-\alpha}{\beta}\right)\mathrm{d}t + \frac{1}{2\beta}\int_{\alpha}^{+\infty} \exp\left(\frac{-t+\alpha}{\beta}\right)\mathrm{d}t \\ &= \frac{1}{2\beta}\lim_{Y\to-\infty}\left[\beta.\exp\left(\frac{t-\alpha}{\beta}\right)\right]_Y^{\alpha} + \lim_{X\to+\infty}\frac{1}{2\beta}\left[-\beta.\exp\left(\frac{\alpha-t}{\beta}\right)\right]_{\alpha}^{X} \\ &= \lim_{Y\to-\infty}\frac{1}{2}\left(1-\exp\left(\frac{Y-\alpha}{\beta}\right)\right) + \lim_{X\to+\infty}\frac{1}{2}\left(1-\exp\left(\frac{\alpha-X}{\beta}\right)\right)\end{aligned}$$

$$\int_{-\infty}^{+\infty} f(t)\mathrm{d}t = \frac{1}{2}.(1-0) + \frac{1}{2}.(1-0) = 1$$

   Ce qui achève de prouver que f est bien une densité de probabilité d'une variable aléatoire réelle.
2. La fonction de répartition, notée Ψ, de la loi $\mathcal{L}(0,1)$, est définie par :

$$\forall x \in \mathbb{R}, \quad \Psi(x) = \int_{-\infty}^{x} \frac{1}{2}e^{-|t|}\,\mathrm{d}t$$

   On distingue deux cas de figure, suivant le signe de x :
   - Pour tout $x \in]-\infty; 0]$:

$$\Psi(x) = \int_{-\infty}^{x} \frac{1}{2}e^{t}\,\mathrm{d}t = \lim_{Y\to-\infty}\frac{1}{2}\left(e^x - e^Y\right) = \frac{1}{2}e^x$$

- Pour tout réel $x \in]0;+\infty[$:

$$\Psi(x) = \int_{-\infty}^{0} \frac{1}{2}e^{t}\mathrm{d}t + \int_{0}^{x} \frac{1}{2}e^{-t}\mathrm{d}t = \lim_{Y\to-\infty} \frac{1}{2}\Big(e^{0} - e^{Y}\Big) + \frac{1}{2}\Big[-e^{-t}\Big]_{0}^{x} = \frac{1}{2} - \frac{1}{2}e^{-x} + \frac{1}{2} = 1 - \frac{1}{2}e^{-x}$$

3. On suppose que X suit la loi $\mathcal{L}(0,1)$.

a) On obtient la loi de $Y = \beta X + \alpha$ par le calcul de sa fonction de répartition :

$$\forall x \in \mathbb{R}, \quad F_Y(x) = \mathbb{P}(Y \leqslant x) = \mathbb{P}(\beta X + \alpha \leqslant x) = \mathbb{P}\Big(X \leqslant \frac{x-\alpha}{\beta}\Big) \qquad \text{car } \beta > 0$$

Soit : $\quad \forall x \in \mathbb{R}, \quad F_Y(x) = \Psi\Big(\frac{x-\alpha}{\beta}\Big)$.

Comme la densité d'une loi $\mathcal{L}(0,1)$ est continue sur tout $\mathbb{R}$, la fonction de répartition associée Ψ est de classe $\mathcal{C}^1$ sur tout $\mathbb{R}$: par composition avec la fonction affine $x \mapsto \dfrac{x-\alpha}{\beta}$, F_Y est bien de classe $\mathcal{C}^1$ sur $\mathbb{R}$, donc Y est une variable à densité, dont une densité est définie par dérivation de la composée F_Y :

$$\forall x \in \mathbb{R}, \quad f_Y(x) = \frac{1}{\beta}\Psi'\Big(\frac{x-\alpha}{\beta}\Big) = \frac{1}{\beta}f_X\Big(\frac{x-\alpha}{\beta}\Big) = \frac{1}{2\beta}\exp\left(-\frac{|x-\alpha|}{\beta}\right)$$

ce qui correspond bien à la densité d'une loi $\mathcal{L}(\alpha,\beta)$.

b) La fonction de répartition de la loi $\mathcal{L}(\alpha,\beta)$ est donc celle de $Y = \beta X + \alpha$, où X suit la loi $\mathcal{L}(0,1)$, de sorte que :

$$\forall x \in \mathbb{R}, \quad F_Y(x) = \Psi\left(\frac{x-\alpha}{\beta}\right) = \begin{cases} \dfrac{1}{2}\exp\left(\dfrac{x-\alpha}{\beta}\right) & \text{si } \dfrac{x-\alpha}{\beta} \leqslant 0 \iff \boxed{x \leqslant \alpha} \\ 1 - \dfrac{1}{2}\exp\left(-\dfrac{x-\alpha}{\beta}\right) & \text{si } \dfrac{x-\alpha}{\beta} > 0 \iff \boxed{x > \alpha} \end{cases}$$

4. *Espérance et variance.*

a) On suppose que X suit la loi $\mathcal{L}(0,1)$. Cette variable aléatoire admet une espérance si et seulement si l'intégrale $\displaystyle\int_{-\infty}^{+\infty} \frac{x}{2}e^{-|x|}\mathrm{d}x$, est absolument convergente.

En remarquant ici que la fonction $g : \ x \mapsto \dfrac{x}{2}e^{-|x|}$ est impaire :

$\forall x \in \mathbb{R}, \ g(-x) = \dfrac{-x}{2}e^{-|-x|} = -\dfrac{x}{2}e^{-|x|} = g(x)$, et positive sur $\mathbb{R}_+$, il suffit alors de prouver l'absolue convergence de l'intégrale $\displaystyle\int_{0}^{+\infty} \frac{x}{2}e^{-|x|}\mathrm{d}x = \frac{1}{2}\int_{0}^{+\infty} xe^{-x}\mathrm{d}x$.

Or $\displaystyle\int_{0}^{+\infty} xe^{-x}\mathrm{d}x$ est l'intégrale définissant l'espérance d'une variable aléatoire T suivant la loi exponentielle de paramètre 1 : elle est convergente (et vaut 1), donc X admet une espérance qui vaut :

$$E(X) = \int_{-\infty}^{0} xf_X(x)\mathrm{d}x + \int_{0}^{+\infty} xf_X(x)\mathrm{d}x = -\int_{0}^{+\infty} xf_X(x)\mathrm{d}x + \int_{0}^{+\infty} xf_X(x)\mathrm{d}x = 0$$

D'après le théorème de transfert : la variable aléatoire X admet un moment d'ordre 2 si et seulement si l'intégrale $\displaystyle\int_{-\infty}^{+\infty} \frac{x^2}{2}e^{-|x|}\mathrm{d}x$ est absolument convergente.

Comme la fonction $h : \ x \mapsto \dfrac{x^2}{2}e^{-|x|}$ est positive et paire, il suffit de prouver la convergence

simple de $\displaystyle\int_0^{+\infty} \frac{x^2}{2}e^{-x}\mathrm{d}x$. On reconnaît à un facteur $\frac{1}{2}$ près, le moment d'ordre 2 de la variable T précédemment introduite, qui vaut d'après la formule de Koenig-Huygens :

$$E(T^2) = V(T) + E(T)^2 = 1 + 1^2 = 2.$$

Ainsi, X admet un moment d'ordre 2 qui vaut :

$$E(X^2) = 2\int_0^{+\infty} \frac{x^2}{2}e^{-x}\mathrm{d}x = \int_0^{+\infty} x^2 e^{-x}\mathrm{d}x = E(T^2) = 2$$

La variable aléatoire X admet donc une variance qui vaut : $V(X) = E(X^2) - E(X)^2 = 2 - 0 = 2$.

b) D'après ce qui précède, la variable aléatoire $Y = \beta X + \alpha$ suit la loi $\mathcal{L}(\alpha, \beta)$ et admet alors une espérance et une variance données par :

$$E(Y) = \beta E(X) + \alpha = \alpha \quad \text{et} \quad V(Y) = \beta^2 V(X) = 2\beta^2$$

par linéarité de l'espérance, et d'après les propriétés de la variance.

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 $\frac{1}{2}$, indépendante de U.

a) Pour tout réel x, le calcul de $P(X \leqslant x)$ se fait via la formule des probabilités totales, appliquée avec le s.c.e. $\big([V=0],[V=1]\big)$:

$$\begin{aligned}
\mathbb{P}(X \leqslant x) &= \mathbb{P}\big([V=0] \cap [(2V-1)U \leqslant x]\big) + \mathbb{P}\big([V=1] \cap [(2V-1)U \leqslant x]\big) \\
&= \mathbb{P}\big([V=0] \cap [-U \leqslant x]\big) + \mathbb{P}\big([V=1] \cap [U \leqslant x]\big) \\
&= \mathbb{P}(V=0) \times \mathbb{P}(U \geqslant -x) + \mathbb{P}(V=1) \times \mathbb{P}(U \leqslant x) \quad \text{par indépendance de } U \text{ et } V \\
&= \frac{1}{2}\big(1 - F_U(-x)\big) + \frac{1}{2}F_U(x)
\end{aligned}$$

On distingue alors deux cas, suivant le signe de x :

$$F_X(x) = \begin{cases} \frac{1}{2}(1-0) + \frac{1}{2}(1-e^{-x}) = 1 - \frac{1}{2}e^{-x} & \text{si } x > 0 \\ \frac{1}{2}(1-1+e^{x}) + 0 = \frac{1}{2}e^{x} & \text{si } x \leqslant 0 \end{cases}$$

On retrouve exactement l'expression de Ψ, ce qui permet de conclure que X suit bien la loi $\mathcal{L}(0,1)$.

b) Le principe du script ci-dessous est alors simple : on simule une réalisation de U et une de V via les fonctions de simulations usuelles connues en `Scilab` : le calcul de $X = (2V-1)U$ correspond alors à une simulation de la loi $\mathcal{L}(0,1)$, et celui de $Y = \beta X + \alpha$ correspond à la simulation de la loi $\mathcal{L}(\alpha, \beta)$.

```
function r = Laplace(alpha,beta)
    if rand() <= 1/2 then    // simulation de V
        V = 1
    else
        V = 0
    end
    X = (2*V-1) * grand(1,1,"exp",1)
    r = beta * X + alpha
endfunction
```

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 $\mathbb{R}$:

$$e^{-\varepsilon}\mathbb{P}\big([X \in I]\big) \leqslant P\big([Y \in I]\big) \leqslant e^{\varepsilon}\mathbb{P}\big([X \in I]\big)$$

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) Si (X, Y) est ε-différentiel, alors pour tout intervalle I de $\mathbb{R}$:

   $$e^{-\varepsilon}P([X \in I]) \leqslant P([Y \in I]) \iff P([X \in I]) \leqslant e^{\varepsilon}P([Y \in I]) \quad \text{multiplication par } e^{\varepsilon} > 0$$

   et :

   $$P([Y \in I]) \leqslant e^{\varepsilon}P([X \in I]) \iff e^{-\varepsilon}P([Y \in I]) \leqslant P([X \in I]) \quad \text{division par } e^{\varepsilon} > 0$$

   donc :

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

   ce qui prouve que le couple (Y, X) est aussi ε-différentiel.

   b) Supposons que (X, Y) est ε-différentiel, et que (Y, Z) est ε'-différentiel, alors pour tout intervalle I de $\mathbb{R}$:

   $$e^{-\varepsilon}P([X \in I]) \leqslant P([Y \in I]) \Longrightarrow e^{-(\varepsilon+\varepsilon')}P([X \in I]) \leqslant e^{-\varepsilon'}P([Y \in I]) \leqslant P([Z \in I])$$

   et par ailleurs :

   $$P([Y \in I]) \leqslant e^{\varepsilon}P([X \in I]) \Longrightarrow e^{\varepsilon'}P([Y \in I]) \leqslant e^{\varepsilon+\varepsilon'}P([X \in I]), \quad \text{or } P([Z \in I]) \leqslant e^{\varepsilon'}P([Y \in I])$$

   Par transitivité de l'inégalité, les deux inégalités précédentes donnent bien :

   $$e^{-(\varepsilon+\varepsilon')}P([X \in I]) \leqslant P([Z \in I]) \leqslant e^{\varepsilon+\varepsilon'}P([X \in I])$$

   ce qui prouve que le couple (X, Z) est $(\varepsilon + \varepsilon')$-différentiel.

7. Soit (X, Y) un couple de variables aléatoires réelles discrètes.
   On suppose que $X(\Omega) \cup Y(\Omega) = \{z_n |\ n \in J\}$, où J est un sous-ensemble non vide de $\mathbb{N}$.
   Montrons par double implication l'équivalence :

   $$(X, Y) \text{ est } \varepsilon\text{-différentiel} \quad \iff \quad \forall n \in J, \quad e^{-\varepsilon}P([X = z_n]) \leqslant P([Y = z_n]) \leqslant e^{\varepsilon}P([X = z_n])$$

   - Si (X, Y) est ε-différentiel : l'ensemble $X(\Omega) \cup Y(\Omega)$ étant un ensemble *discret*, pour tout entier $n \in J$, il existe un intervalle réel I_n qui ne contient que z_n (par exemple du type : $]z_n - \alpha; z_n + \alpha[$ avec $\alpha > 0$ suffisament petit), et alors :

     $$e^{-\varepsilon}P([X \in I_n]) \leqslant P([Y \in I_n]) \leqslant e^{\varepsilon}P([X \in I_n]) \iff e^{-\varepsilon}P([X = z_n]) \leqslant P([Y = z_n]) \leqslant e^{\varepsilon}P([X = z_n])$$

   - Réciproquement, si pour tout $n \in J$, $e^{-\varepsilon}P([X = z_n]) \leqslant P([Y = z_n]) \leqslant e^{\varepsilon}P([X = z_n])$, alors pour tout intervalle I de $\mathbb{R}$: $P([X \in I]) = \sum_{n \in J \text{ tq } z_n \in I} P([X = z_n])$, et par sommation de l'inégalité précédente pour tout $n \in J$ tel que $z_n \in I$, on obtient :

     $$e^{-\varepsilon} \sum_{n \in J \text{ tq } z_n \in I} P([X = z_n]) \leqslant \sum_{n \in J \text{ tq } z_n \in I} P([Y = z_n]) \leqslant e^{\varepsilon} \sum_{n \in J \text{ tq } z_n \in I} P([X = z_n])$$

     ce qui est bien :

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

     Ceci étant valable pour tout intervalle I de $\mathbb{R}$, on en déduit que (X, Y) est bien ε-différentiel.

8. *Premier exemple.*

Dans cette question, on suppose que X suit la loi géométrique de paramètre $\frac{1}{2}$:

$\forall k \in \mathbb{N}^*,\ P([X=k]) = \dfrac{1}{2^{k-1}} \times \dfrac{1}{2} = \dfrac{1}{2^k}$, et Z suit la loi de Bernoulli de paramètre $p \in]0,1[$: $P([Z=1]) = p \quad$ et $\quad P([Z=0]) = 1-p$; on suppose que X et Z sont indépendantes.

On pose $Y = X + Z$.

a) Comme $X(\Omega) = \mathbb{N}^*$ et $Y(\Omega) = \{0,1\}$, alors $Y(\Omega) = \mathbb{N}^*$, et :

$$P([Y=1]) = P([X=1] \cap P([Z=0]) = P([X=1]) \times P([Z=0]) = \frac{1-p}{2}$$

par indépendance de X et Z, tandis que pour tout entier $k \geqslant 2$:

$$[Y=k]) = ([X=k-1] \cap [Z=1]) \cup ([X=k] \cap [Z=0])$$

donc par union disjointe et indépendance de X et Z :

$$\begin{aligned} P([Y=k]) &= P([X=k-1]) \times P([Z=1]) + P([X=k]) \times P([Z=0]) \\ &= \frac{1}{2^{k-1}}p + \frac{1}{2^k}(1-p) = \frac{2p+1-p}{2^k} = \frac{1+p}{2^k} \end{aligned}$$

b) D'après ce qui précède : $\dfrac{P([Y=1])}{P([X=1])} = \dfrac{1-p}{2} \times 2 = 1-p$, qui vérifie bien :

$$1-p \leqslant 1-p = \frac{P([Y=1])}{P([X=1])} \leqslant 1 \leqslant \frac{1}{1-p} \quad \text{puisque } 0 < 1-p < 1$$

Pour tout entier $k \geqslant 2$: $\dfrac{P([Y=k])}{P([X=k])} = \dfrac{1+p}{2^k} \times 2^k = 1+p.$

On a bien : $1-p < 1+p$ puisque $p > 0$, tandis que : $(1+p) \times (1-p) = 1-p^2 < 1$ puisque $0 < p < 1$, ce qui implique : $1+p \leqslant \dfrac{1}{1-p}$, et ainsi on a bien démontré :

$$\forall k \in \mathbb{N}^*, \quad 1-p \leqslant \frac{P([Y=k])}{P([X=k])} \leqslant \frac{1}{1-p}$$

c) La relation précédente se réécrit : $\forall k \in \mathbb{N}^*,\ e^{\ln(1-p)} \leqslant \dfrac{P([Y=k])}{P([X=k])} \leqslant e^{-\ln(1-p)}$, ce qui s'écrit :

$$\forall k \in \mathbb{N}^*, \quad e^{-\varepsilon}P([X=k]) \leqslant P([Y=k]) \leqslant e^{\varepsilon}P([X=k])$$

en posant $\varepsilon = -\ln(1-p) > 0$ puisque $0 < 1-p < 1 \iff \ln(1-p) < 0$.

Étant donné que $Y(\Omega) = \mathbb{N}^*$, et d'après l'équivalence obtenue à la question 7., on en déduit que le couple (X,Y) est $-\ln(1-p)$-différentiel.

d) Lorsque p tend vers 0 : par encadrement dans la relation obtenue en 8.b), $\dfrac{P([Y=k])}{P([X=k])}$ se rapproche de 1, c'est-à-dire que les lois de X et Y sont très proches l'une de l'autre. C'est cohérent avec le fait que la variable de Bernoulli Z a alors une probabilité très forte de prendre la valeur 0, donc Y de prendre la même valeur que X.

Si p se rapproche de 1 : alors Z prend très probablement la valeur 1, et Y est plus proche de $X+1$ que de X, ce qui est cohérent avec le fait que $-\ln(1-p)$ devient d'autant plus grand que p est proche de 1.

9. On suppose que X et Y sont deux variables à densité de densités respectives f et g et de fonctions de répartition F et G.

a) On suppose que pour tout $t \in \mathbb{R}$, $e^{-\varepsilon} f(t) \leqslant g(t) \leqslant e^{\varepsilon} f(t)$.

Les fonctions f et g sont continues sur $\mathbb{R}$ sauf peut-être en un nombre fini de points, et intégrables sur $\mathbb{R}$, donc par croissance de l'intégrale, pour tout intervalle réel I :

$$\int_I e^{-\varepsilon} f(t)\mathrm{d}t \leqslant \int_I g(t)\mathrm{d}t \leqslant \int_I e^{\varepsilon} f(t)\mathrm{d}t \iff e^{-\varepsilon}\int_I f(t)\mathrm{d}t \leqslant \int_I g(t)\mathrm{d}t \leqslant e^{\varepsilon}\int_I f(t)\mathrm{d}t$$

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

et le couple (X, Y) est bien 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. Avec l'intervalle $I = [t; t+h]$, la propriété de ε-différentialité s'écrit :

$$e^{-\varepsilon} P([X \in [t; t+h]]) \leqslant P([Y \in [t; t+h]) \leqslant e^{\varepsilon} P([X \in [t; t+h]])$$

$$\iff e^{-\varepsilon}\big(F(t+h) - F(t)\big) \leqslant G(t+h) - G(t) \leqslant e^{\varepsilon}\big(F(t+h) - F(t)\big)$$

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

Comme t est un point en lequel f et g sont continues, c'est un point en lequel F et G sont de classe $\mathcal{C}^1$, et ainsi : $\lim\limits_{h\to 0^+} \dfrac{F(t+h) - F(t)}{h} = f(t)$, et de même $\lim\limits_{h\to 0^+} \dfrac{G(t+h) - G(t)}{h} = g(t)$.

On peut donc passer à la limite dans la double inégalité précédente, ce qui donne directement :

$$e^{-\varepsilon} f(t) \leqslant g(t) \leqslant e^{\varepsilon} f(t)$$

10. *Deuxième exemple : lois de Cauchy.*

a) La fonction $t \mapsto \dfrac{1}{t^2+1}$ est définie, continue et strictement positive sur $\mathbb{R}$, puisque $t^2 + 1 > 0$ pour tout réel t. Par ailleurs : $\dfrac{1}{t^2+1} \underset{t\to+\infty}{\sim} \dfrac{1}{t^2}$.

Or l'intégrale $\displaystyle\int_1^{+\infty} \frac{1}{t^2}\mathrm{d}t$ est convergente, comme intégrale de Riemann avec $\alpha = 2 > 1$.

Le théorème de comparaison des intégrales de fonctions continues, positives assure alors que l'intégrale $\displaystyle\int_1^{+\infty} \frac{1}{t^2+1}\mathrm{d}t$ est convergente. La continuité de $t \mapsto \dfrac{1}{t^2+1}$, puis la parité de cette fonction sur $\mathbb{R}$, assurent alors que l'intégrale $\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t^2+1}\mathrm{d}t$ est convergente.

On admet que cette intégrale est égale à π.

b) On définit, pour $a > 0$, la fonction f_a sur $\mathbb{R}$ par : $\forall t \in \mathbb{R},\ f_a(t) = \dfrac{a}{\pi(t^2+a^2)}$.

La fonction f_a est continue, strictement positive sur $\mathbb{R}$, il reste donc seulement à démontrer que $\displaystyle\int_{-\infty}^{+\infty} \frac{a}{\pi(t^2+a^2)}\mathrm{d}t = 1$; sous réserve de convergence :

$$\int_{-\infty}^{+\infty} \frac{a}{\pi(t^2+a^2)}\mathrm{d}t = \frac{a}{\pi}\int_{-\infty}^{+\infty} \frac{1}{a^2\Big(\big(\frac{t}{a}\big)^2+1\Big)}\mathrm{d}t = \frac{1}{\pi}\int_{-\infty}^{+\infty} \frac{1}{\big(\frac{t}{a}\big)^2+1}\frac{\mathrm{d}t}{a}$$

Le changement de variable affine : $u = \dfrac{t}{a}$ assure que les intégrales $\dfrac{1}{\pi}\displaystyle\int_{-\infty}^{+\infty} \frac{1}{\big(\frac{t}{a}\big)^2+1}\frac{\mathrm{d}t}{a}$ et $\dfrac{1}{\pi}\displaystyle\int_{-\infty}^{+\infty} \frac{1}{t^2+1}\mathrm{d}t$ sont de même nature, et égales en cas de convergence : c'est bien le cas ici, et elles sont toutes deux égales à 1 d'après la question précédente.

On a ainsi démontré que f_a est une densité de probabilité.

c) On suppose que X et Y sont deux variables aléatoires admettant comme densités respectives f_1 et f_a avec $a > 1$.

Comme $a > 1$, alors pour tout $t \in \mathbb{R}$: $\quad f_a(t) = \dfrac{a}{\pi(t^2+a^2)} \leqslant \dfrac{a}{\pi(t^2+1)} = af_1(t)$.

Par ailleurs : $\quad f_a(t) = \dfrac{a}{\pi(t^2+a^2)} = \dfrac{1}{a\pi} \times \dfrac{1}{\dfrac{t^2}{a^2}+1} \geqslant \dfrac{1}{a\pi} \times \dfrac{1}{t^2+1}$ puisque $\dfrac{t^2}{a^2} < t^2$, étant donné que $a > 1$.

Ainsi : $\quad \forall t \in \mathbb{R}, \quad \dfrac{1}{a}f_1(t) \leqslant f_a(t) \leqslant af_1(t) \iff \forall t \in \mathbb{R}, \quad e^{-\ln(a)}f_1(t) \leqslant f_a(t) \leqslant e^{\ln(a)}f_1(t)$

ce qui, d'après la fonction 9.a), prouve que le couple (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 $\mathbb{R}$ tel que $P([Z \in I]) \neq 0$. Par définition de Z :

$P_{[Z\in I]}(U = 1) = \dfrac{P([Z \in I] \cap [U = 1])}{P([Z \in I])}$, où :

$P([Z \in I] \cap [U = 1]) = P([U = 1]) \times P_{[U=1]}([Z \in I]) = p \times P_{[U=1]}(X \in I) = p \times P(X \in I)$ puisque X et U sont indépendantes.

La formule des probabilités totales, appliquée avec le système complet d'événements $\big([U = 0], [U = 1]\big)$, donne par ailleurs :

$$\begin{aligned} P([Z \in I]) &= P(U = 0) \times P_{[U=0]}(Z \in I) + P(U = 1) \times P_{[U=1]}(Z \in I) \\ &= p \times P_{[U=0]}(Y \in I) + (1-p) \times P_{[U=1]}(X \in I) \\ &= pP([X \in I]) + (1-p)P([Y \in I]) \end{aligned}$$

puisque X et Y sont indépendantes de U. On a bien démontré :

$$P_{[Z\in I]}([U = 1]) = p\frac{P([X \in I])}{pP([X \in I]) + (1-p)P([Y \in I])}$$

Puisque (X, Y) est ε-différentiel :

$$\begin{aligned} & e^{-\varepsilon}P([X \in I]) \leqslant P([Y \in I]) \leqslant e^{\varepsilon}P([X \in I]) \\ \iff & (1-p)e^{-\varepsilon}P([X \in I]) \leqslant (1-p)P([Y \in I]) \leqslant (1-p)e^{\varepsilon}P([X \in I]) \quad \text{car } 1-p > 0 \\ \iff & \big(p + (1-p)e^{-\varepsilon}\big)P([X \in I]) \leqslant pP([X \in I]) + (1-p)P([Y \in I]) \leqslant \big(p + (1-p)e^{\varepsilon}\big)P([X \in I]) \\ \iff & \frac{1}{\big(p + (1-p)e^{-\varepsilon}\big)P([X \in I])} \geqslant \frac{1}{pP([X \in I]) + (1-p)P([Y \in I])} \geqslant \frac{1}{\big(p + (1-p)e^{\varepsilon}\big)P([X \in I])} \\ \implies & \frac{p}{p + (1-p)e^{-\varepsilon}} \geqslant p\frac{P([X \in I])}{pP([X \in I]) + (1-p)P([Y \in I])} \geqslant \frac{p}{p + (1-p)e^{\varepsilon}} \quad \text{car } pP([X \in I]) \geqslant 0 \end{aligned}$$

Soit : $\quad \dfrac{p}{p + (1-p)e^{\varepsilon}} \leqslant P_{[Z\in I]}([U = 1]) \leqslant \dfrac{p}{p + (1-p)e^{-\varepsilon}}$

b) Si ε est proche de zéro, alors $\dfrac{p}{p+(1-p)e^{\varepsilon}}$ et $\dfrac{p}{p+(1-p)e^{-\varepsilon}}$ sont tous les deux très proches de $\dfrac{p}{p+(1-p)} = p$, donc par encadrement, $P_{[Z\in I]}(U=1)$ est très proche de $P(U=1)$.
Cela signifie que dans ce cas, le fait de disposer d'une information sur la valeur de Z ne change pas notablement le paramètre de la loi de U, donc pas non plus 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 diffèrent 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 $a = (a_1, a_2, \ldots, a_n)$, 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 $a_2, \ldots, a_n$ sont connus et on cherche à protéger a_1.

a) Si on choisit une valeur au hasard dans $[\![0, d]\!]$, on a évidemment une probabilité $\dfrac{1}{d+1}$ d'obtenir la bonne valeur de a_1.

b) Dans cette question, $q(a_1, \ldots, a_n) = \displaystyle\sum_{i=1}^{n} a_i$.

Si $q(a)$ est publique, $a_2, \ldots, a_n$ étant connus, alors évidemment $a_1 = q(a) - \displaystyle\sum_{i=2}^{n} a_i$ est connu lui aussi !

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 $a_2, \ldots, a_n$ sont connus, que l'on cherche à protéger a_1 et que :

- Le public connaît des intervalles $I_0, \ldots, I_d$ disjoints de réunion $\mathbb{R}$ tels qu'avec les valeurs fixées de $a_2, \ldots, a_n$, 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-confidentitalité 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 $j \in [\![0, d]\!]$, $Y_j = X_{(j, a_2, \ldots, a_n)}$. On suppose que A_1 et Y_j sont indépendantes pour tout $j \in D$.
- la variable aléatoire réelle R par :

$$\forall \omega \in \Omega, \text{ si } A_1(\omega) = j \text{ alors on détermine l'unique } k \text{ tel que } Y_j(\omega) \in I_k \text{ et on pose } R(\omega) = k.$$

- $\theta = P([R = A_1]))$.

a) L'événement $[R = A_1]$ se décompose sous la forme de la réunion disjointe :

$$[R = A_1] = \bigcup_{j=0}^{d} \big([R = j] \cap [A_1 = j]\big) = \bigcup_{j=0}^{d} \big([Y_j \in I_j] \cap [A_1 = j]\big)$$

et ainsi on a bien : $\theta = P([R = A_1]) = \sum_{j=0}^{d} P([Y_j \in I_j] \cap [A_1 = j])$.

b) L'énoncé stipule bien que A_1 est indépendante de chacune des variables aléatoires Y_j, et que A_1 suit en fait la loi uniforme sur $[\![0, d]\!]$, donc :

$$\theta = \sum_{j=0}^{d} P([Y_j \in I_j]) \times P(A_1 = j) = \frac{1}{d+1} \sum_{j=0}^{d} P([Y_j \in I_j]).$$

c) Au vu des définitions : pour tout entier $j \in [\![1, d]\!]$, les variables aléatoires $Y_j = X_{(j,a_2,\dots,a_n)}$ et $Y_0 = X_{(0,a_2,\dots,a_n)}$ forment un couple ε-différentiel, puisque $(j, a_2, \dots, a_n)$ et $(0, a_2, \dots, a_n)$ sont voisins (ils ne diffèrent que par leur première coordonnée). Par conséquent :

$$\forall j \in [\![1, d]\!], \quad e^{-\varepsilon} P([Y_j \in I_j]) \leqslant P([Y_0 \in I_j]) \leqslant e^{\varepsilon} P([Y_j \in I_j])$$

La sommation membre à membre de l'inégalité de droite pour j variant de 1 à d donne :

$$\sum_{j=1}^{d} P([Y_0 \in I_j]) \leqslant e^{\varepsilon} \sum_{j=1}^{d} P([Y_j \in I_j]) \iff \sum_{j=0}^{d} P([Y_j \in I_j]) \leqslant e^{\varepsilon} \sum_{j=1}^{d} P([Y_0 \in I_0]) + P([Y_0 \in I_0])$$

où : $\sum_{j=1}^{d} P([Y_0 \in I_j]) = 1 - P([Y_0 \in I_0])$; en effet, puisque $(I_0, I_1, \dots, I_d)$ forment une partition de $\mathbb{R}$, alors les événements $\big([Y_0 \in I_j]\big)_{0 \leqslant j \leqslant n}$ forment un système complet d'événements.

D'autre part, d'après b) :

$\sum_{j=0}^{d} P([Y_j \in I_j]) = \theta(d+1) \iff \sum_{j=1}^{d} P([Y_j \in I_j]) = \theta(d+1) - P([Y_0 \in I_0])$; tout ceci permet de réécrire l'inégalité précédente sous la forme :

$$\begin{aligned} &\theta(d+1) \leqslant e^{\varepsilon}\big(1 - P([Y_0 \in I_0])\big) + P([Y_0 \in I_0]) \\ \iff &\theta(d+1) \leqslant e^{\varepsilon} - (e^{\varepsilon} - 1)P([Y_0 \in I_0]) \\ \iff &\theta \leqslant \frac{1}{d+1}\Big(e^{\varepsilon} - (e^{\varepsilon} - 1)P([Y_0 \in I_0])\Big) \end{aligned}$$

Il reste à remarquer que puisque $\varepsilon > 0$, alors $e^{\varepsilon} > 1 \iff (e^{\varepsilon} - 1) > 0$ et $P([Y_0 \in I_0]) \geqslant 0$ (c'est une probabilité), et ainsi on a bien :

$$\theta \leqslant \frac{1}{d+1}\Big(e^{\varepsilon} - (e^{\varepsilon} - 1)P([Y_0 \in I_0])\Big) \leqslant \frac{e^{\varepsilon}}{d+1}$$

d) On pose $\rho = \dfrac{1}{d+1}$ et $\tau = \dfrac{\theta - \rho}{\rho}$.

Au vu de ces définitions et du résultat précédent :

$$\tau = \big(\theta - \frac{1}{d+1}\big) \times (d+1) = \theta(d+1) - 1 \leqslant e^{\varepsilon} - 1$$

Le réel τ représente l'erreur, la différence relative entre la probabilité d'obtenir la bonne valeur de a_1 par un choix au hasard avec équiprobabilité dans $[\![0,d]\!]$ (c'est ρ), et la probabilité d'obtenir la bonne valeur de a_1 par la méthode de confidentialité (c'est θ). Lorsque ε est proche de 0, $e^{\varepsilon}-1$ aussi, et τ est alors lui-même proche de 0 : la méthode de confidentialité n'apporte donc pas, dans ce cas, d'avantage particulier par rapport au choix aléatoire, équiprobable envisagé dès le début !

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 paramètres $(0,\beta)$.

a) Des calculs analogues à ceux pratiqués dans la partie I, question 3. donnent :
$E(X_a) = q(a) + E(Y) = q(a) + 0 = q(a)$, tandis qu'une densité f_a de X_a est définie sur $\mathbb{R}$ par :

$$\forall x \in \mathbb{R}, \quad f_a(x) = f_Y\big(x - q(a)\big) = \frac{1}{2\beta}\exp\Big(-\frac{|x-q(a)|}{\beta}\Big)$$

(la variable aléatoire X_a suit en fait la loi de Laplace $\mathcal{L}\big(q(a),\beta\big)$.)

b) Pour tout réel t et tout couple $(a,b) \in \mathcal{V}$: $\dfrac{f_a(t)}{f_b(t)} = \exp\Big(\dfrac{|t-q(b)|-|t-q(a)|}{\beta}\Big)$, où :
$|t-q(b)| = |t-q(a)+q(a)-q(b)| \leqslant |t-q(a)| + |q(a)-q(b)|$ d'après l'inégalité triangulaire, d'où :
$|t-q(a)| - |t-q(b)| \leqslant |q(a)-q(b)| \leqslant \max\limits_{(a,b)\in\mathcal{V}} |q(a)-q(b)| = \delta$.
Comme $\beta > 0$, la stricte croissante de l'exponentielle sur $\mathbb{R}$ permet d'en déduire :

$$\forall t \in \mathbb{R},\ \exp\Big(\frac{|t-q(a)|-|t-q(b)|}{\beta}\Big) \leqslant \exp\Big(\frac{\delta}{\beta}\Big) \iff \frac{f_a(t)}{f_b(t)} \leqslant \exp\Big(\frac{\delta}{\beta}\Big)$$

ce qui donne bien :

$$\forall t \in \mathbb{R},\ \forall (a,b) \in \mathcal{V}, \quad f_a(t) \leqslant \exp\Big(\frac{\delta}{\beta}\Big) f_b(t).$$

Les éléments a et b jouent ici des rôles symétriques, donc on a aussi :
$\forall t \in \mathbb{R},\ f_b(t) \leqslant \exp\Big(\dfrac{\delta}{\beta}\Big) f_a(t) \iff \exp\Big(-\dfrac{\delta}{\beta}\Big) f_b(t) \leqslant f_a(t)$, donc :

$$\forall (a,b) \in \mathcal{V},\ \forall t \in \mathbb{R}, \quad \exp\Big(-\frac{\delta}{\beta}\Big) f_b(t) \leqslant f_a(t) \leqslant \exp\Big(\frac{\delta}{\beta}\Big) f_b(t)$$

ce qui implique bien, d'après la question 9., que le couple (X_a, X_b) est $\dfrac{\delta}{\beta}$-différentiel.

c) Il suffit alors de choisir $\boxed{\beta = \dfrac{\delta}{\varepsilon}}$ pour que pour tout (a,b) de $\mathcal{V}$, le couple (X_a, X_b) soit ε-différentiel.

Comme pour tout $a \in D^n$, $E(X_a) = q(a)$, on a bien dans ce cas un procédé de ε-confidentialité de D^n pour q.

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

a) Soit (a,b) un couple de $\mathcal{V}$: alors $q(a) - q(b) = \sum\limits_{k=1}^{n} (a_i - b_i)$, où dans cette somme, tous les termes sont nuls sauf 1, puisque a et b ne diffèrent que d'une composante (ils sont voisins).
La valeur maximale $|a_i - b_i|$ est égale à d puisque a_i et b_i appartiennent à $[\![0,d]\!]$, donc pour cette définition de l'application q :

$$\delta = d$$

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 $X_a < \dfrac{1}{2}$, 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) Il suffit de lire correctement la définition précédente ! La variable aléatoire Z_a publiée vérifie, pour $\omega \in \Omega$:

- $Z_a(\omega) = 0$ si $X_a(\omega) < \dfrac{1}{2}$;
- s'il existe $k \in [\![1, nd-1]\!]$ tel que $k + \frac{1}{2} \leqslant X_a(\omega) + \frac{1}{2} < k + \frac{1}{2} \iff k \leqslant X_a(\omega) + \frac{1}{2} < k+1$, alors $\lfloor X_a(\omega) + \frac{1}{2} \rfloor = k = Z_a(\omega)$.

  Cette relation est donc vraie pour tout $\omega \in \Omega$ tel que $X_a(\omega) \in \bigcup_{k=1}^{nd-1} [k-\frac{1}{2}; k+\frac{1}{2}[= [\frac{1}{2}; nd-\frac{1}{2}[$.
- sinon, $X_a(\omega) \geqslant nd - \frac{1}{2}$ et alors $Z_a(\omega) = nd$.

c) Le script utilise la fonction `Laplace` écrite à la fin de la partie I :

```
d = input("Saisir d : ")
n = input("Saisir n : ")
eps = input("Saisir epsilon : ")
beta = d/eps
a = grand(1,n,"uin",0,d)
q = sum(a)
Y = Laplace(0,beta)
Xa = q + Y
if Xa < 1/2 then
    Za = 0
elseif Xa < n*d-1/2 then
    Za = floor(Xa + 1/2)
else
    Za = nd
end
disp(q)
disp(Za)
```

d) Pour $n = 1000$, $d = 4$ et ε choisi par l'utilisateur : la valeur moyenne de $\dfrac{|Z_a - q(a)|}{q(a)}$ est obtenue en simulant un grand nombre de fois Z_a, et en calculant la moyenne des valeurs des valeurs de $\dfrac{|Z_a - q(a)|}{q(a)}$ obtenues, qu'on aura enregistrées dans un vecteur de même taille que l'échantillon de la simulation.

On reprend donc le script précédent en le modifiant et le complétant :

```
n = 1000
d = 4
eps = input("Saisir epsilon : ")
beta = d/eps
N = 100000
T = zeros(1,N)
for i = 1:N
    a = grand(1,n,"uin",0,d)
    q = sum(a)
```

```
10          Y = Laplace(0,beta)
11          Xa = q + Y
12          if Xa < 1/2 then
13              Za = 0
14          elseif Xa < n*d-1/2 then
15              Za = floor(Xa + 1/2)
16          else
17              Za = nd
18          end
19          T(i) = abs(Za - q)/q
20      end
21      disp(mean(T))
```

L'exécution du script pour chacune des valeurs de ε du tableau de l'énoncé, donne bien des pourcentages du même ordre que ceux de la deuxième ligne de ce tableau.

★★★ FIN DU SUJET ★★★