Arbitrages statistiques dans l'apprentissage automatique confidentiel.

Rapport de stage

Alexi Canesse

Sous la supervision d'Aurélien Garivier, Professeur,
UMPA et École Normale Supérieure de Lyon

Stage de recherche effectué dans le cadre de la
L3 informatique fondamental de l'ÉNS de Lyon

Département informatique
École Normale Supérieur de Lyon
France

12 juillet 2022

Table des matières

1 Introduction

TODO

1.1 Présentation du problème

TODO

1.2 Définitions

La *differential privacy* [**10.1007/11681878_14**] quantifie la perte de confidentialité subit par un individue en étant dans une base de donné.

Définition 1.2.0.1 : *Bases de donnée voisines*
On dit que deux bases de données x et y sont voisines et on note $||x - y||_1 \leq 1$ si elles diffèrent sur au plus une entrée *ie* la distance de HAMMING qui les sépare et majorée par 1.

Définition 1.2.0.2 : *Differential privacy*
On dit qu'un mécanisme aléatoire $\mathcal{M} : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ est (ε, δ)-***differentially private*** si pour tout $\mathcal{S} \subset \mathcal{T}$ mesurable,

$$\forall x, y \in \mathcal{X}^{(\mathbb{N})} \quad ||x - y|| \leq 1 \quad \Rightarrow \quad \mathbb{P}(\mathcal{M}(x) \in \mathcal{S}) \leq \exp(\varepsilon)\mathbb{P}(\mathcal{M}(y) \in \mathcal{S}) + \delta$$

De plus, si $\delta = 0$, on dit que $\mathcal{M}$ est ε-***differentially private***.

1.3 Contribution

TODO

2 Méthode des histogrammes

2.1 AboveThreshold

Répondre à de nombreuses requête est coûteux en confidentialité. Utiliser à algorithme naïf tel que le mécanisme de LAPLACE [**10.1007/11681878_14**] ne permet pas de répondre à de nombreuses requêtes avec une bonne précision tout en préservant un bon niveau de confidentialité (ε doit être petit). Dans certains cas nous ne sommes néanmoins pas intéressé par les réponses numériques, mais uniquement intéressé par le fait qu'une réponse dépasse ou non un seuil définit. Nous allons voir que `AboveThreshold` [**dwork2014the**] permet cela tout en ne payant que pour les requêtes qui dépassent le seuil.

```
AboveThreshold(database, queries, threshold, epsilon){
    Assert("les requêtes sont toutes de sensibilité 1");
    result = 0;
    noisyThreshold = threshold + Lap(2/epsilon);
    for(querie in queries){
        nu = Lap(4/epsilon);
        if(querie(D) + nu > noisyThreshold)
            return result;
        else
            ++result;
    }
    return -1;
}
```

L'algorithme venant d'être décrit renvoie l'indice de la première requête à dépasser le seuil si une telle requête existe. C'est une version adaptée de l'algorithme initialement décrit par DWORK et ROTH dans [**dwork2014the**]. Icelui a du sens d'un point de vue informatique mais rend le formalisme mathématiques compliqué (les auteurs eux-même tombent dans ce travers) et nous n'utiliseront pas les légers avantages de leur version.

Théorème 2.1.0.1 :
Pour tout ensemble de requêtes $Q \in \left(\mathcal{X}^{(\mathbb{N})} \to \mathcal{T}\right)^{\mathbb{N}}$ de sensibilité 1, tout seuil $T \in \mathbb{R}$, tout $\varepsilon > 0$, $M : x \in \mathcal{X}^{(\mathbb{N})} \mapsto$ `AboveThreshold(x, Q, T, epsilon)` est *ε-differentially private.*

Remarque : *La démonstration est une réécriture de celle du livre de référence [**dwork2014the**]. Une réécriture était nécessaire car cette démonstration présente de nombreux points limites en terme de rigueur mathématiques et de detail pas suffisant sur certains points non triviaux.*

Démonstration :
Soit $D, D' \in \mathcal{X}^{(\mathbb{N})}$ tels que $||D - D'|| \leq 1$, $\{f_i\}_i = Q \in \left(\mathcal{X}^{(\mathbb{N})} \to \mathcal{T}\right)^{\mathbb{N}}$ un ensemble de requêtes de sensibilité 1, $T \in \mathbb{R}$ un seuil, et $\varepsilon > 0$. On pose alors A la variable aléatoire `AboveThreshold(D, Q, T, epsilon)` et A' la variable aléatoire `AboveThreshold(D', Q, T, epsilon)`.

Soit alors $k \in \mathbb{N}$. *Montrons que* $\mathbb{P}(A = k) \leq \exp(\varepsilon)\mathbb{P}(A' = k)$. En reprenant les notations de l'algorithme [2.1], on fixe les éléments $(\nu_i)_{i<k}$ (qui suivent une loi de LAPLACE de paramètre $4/\varepsilon$).

On pose alors

$$\begin{cases} g_k &= \max_{i<k} \{f_i(D) + \nu_i\} \\ g'_k &= \max_{i<k} \{f_i(D') + \nu_i\} \end{cases}$$

Ces grandeurs représente la valeur plus grande comparée au seuil bruité avant l'indice k dans le cas de l'execution sur D et de l'execution sur D'. Les probabilité qui suivent seront prisent sur les deux variables aléatoires non fixées ν_k et $\hat{T}$ qui est la valeur du seuil bruitée. On pose enfin, pour tout $i \in \mathbb{N}$,

$$\begin{cases} y_i &= f_i(D) \\ y'_i &= f_i(D') \end{cases}$$

On note alors que, en notant l_2 la densité de la loi de LAPLACE de paramètre $2/\varepsilon$ et l_4 celle de paramètre $4/\varepsilon$,

$$\begin{aligned} \mathbb{P}(A = k) &= \mathbb{P}(\hat{T} \in]g_k, y_k + \nu_k]) \\ &= \int_{\mathbb{R}} \mathbb{P}(\hat{T} \in]g_k, y_k + \nu]) l_4(\nu) \mathrm{d}\nu \\ &= \int_{\mathbb{R}} \int_{g_k - T}^{y_k + \nu - T} l_2(t) l_4(\nu) \mathrm{d}t \mathrm{d}\nu \end{aligned}$$

On effectue alors un premier changement de variable affine

$$\hat{t} = t + g_k - g'_k$$

On obtient donc

$$\begin{aligned} \mathbb{P}(A = k) &= \int_{\mathbb{R}} \int_{g_k - T}^{y_k + \nu - T} l_2(\hat{t} - g_k + g'_k) l_4(\nu) \mathrm{d}t \mathrm{d}\nu \\ &= \int_{\mathbb{R}} \int_{g'_k - T}^{y_k + \nu - g_k + g'_k - T} l_2(\hat{t}) l_4(\nu) \mathrm{d}t \mathrm{d}\nu \end{aligned}$$

Il est alors temps de faire un second changement de variable affine

$$\hat{\nu} = \nu + g_k - g'_k + y'_k - y_k$$

Ainsi,

$$\begin{aligned}
\mathbb{P}(A = k) &= \int_{\mathbb{R}} \int_{g'_k - T}^{y_k + \nu - g_k + g'_k - T} l_2(\hat{t}) l_4(\hat{\nu} - g_k + g'_k - y'_k + y_k) \mathrm{d}\hat{t}\mathrm{d}\nu \\
&= \int_{\mathbb{R}} \int_{g'_k - T}^{y_k + \nu - g_k + g'_k + g_k - g'_k + y'_k - y_k - T} l_2(\hat{t}) l_4(\hat{\nu}) \mathrm{d}t\mathrm{d}\nu \\
&= \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} l_2(\hat{t}) l_4(\hat{\nu}) \mathrm{d}t\mathrm{d}\nu
\end{aligned}$$

Par définition de l_2 et l_4 nous avons donc

$$\mathbb{P}(A = k) = \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} \exp\left(-\frac{|\hat{t}|\varepsilon}{2}\right) \exp\left(-\frac{|\hat{\nu}|\varepsilon}{4}\right) \mathrm{d}t\mathrm{d}\nu$$

L'inégalité triangulaire assure alors que

$$\begin{aligned}
\mathbb{P}(A = k) &\leq \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} \exp\left(\frac{|\hat{t} - t|\varepsilon}{2}\right) \exp\left(-\frac{|t|\varepsilon}{2}\right) \exp\left(\frac{|\hat{\nu} - \nu|\varepsilon}{4}\right) \exp\left(-\frac{|\nu|\varepsilon}{4}\right) \mathrm{d}t\mathrm{d}\nu \\
&= \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} \exp\left(\frac{|g_k - g'_k|\varepsilon}{2}\right) \exp\left(-\frac{|t|\varepsilon}{2}\right) \exp\left(\frac{|g_k - g'_k + y'_k - y_k|\varepsilon}{4}\right) \exp\left(-\frac{|\nu|\varepsilon}{4}\right) \mathrm{d}t\mathrm{d}\nu
\end{aligned}$$

Les requêtes étant de sensibilité 1, nous avons

$$\begin{cases} 2 & \geq |g_k - g'_k| + |y'_k - y_k| \quad \geq |g_k - g'_k + y'_k - y_k| \\ 1 & = |g_k - g'_k| \end{cases}$$

Enfin, la croissance de l'intégrale assure que

$$\begin{aligned}
\mathbb{P}(A = k) &\leq \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} \exp\left(\frac{\varepsilon}{2}\right) \exp\left(-\frac{|t|\varepsilon}{2}\right) \exp\left(\frac{\varepsilon}{2}\right) \exp\left(-\frac{|\nu|\varepsilon}{4}\right) \mathrm{d}t\mathrm{d}\nu \\
&= \exp\left(\frac{2\varepsilon}{2}\right) \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} \exp\left(-\frac{|t|\varepsilon}{2}\right) \exp\left(-\frac{|\nu|\varepsilon}{4}\right) \mathrm{d}t\mathrm{d}\nu \\
&= \exp(\varepsilon) \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} l_2(t) l_4(\nu) \mathrm{d}t\mathrm{d}\nu \\
&= \exp(\varepsilon) \int_{\mathbb{R}} \mathbb{P}(\hat{T} \in]g'_k, y'_k + \nu]) l_4(\nu) \mathrm{d}\nu \\
&= \exp(\varepsilon)\, \mathbb{P}(\hat{T} \in]g'_k, y'_k + \nu_k]) \\
&= \exp(\varepsilon)\, \mathbb{P}(A' = k)
\end{aligned}$$

2.2 La méthode des histogramme

2.2.1 Présentation de la méthode des histogrammes

La méthode des histogramme est une méthode que nous avons proposé durant ce stage. Il s'agit d'une instanciation particulière de `AboveThreshold` permettant de calculer l'ensemble des déciles (ou n'importe quel quantiles). Une transformation affine permet d'obtenir la réponse finale à partir de la réponse du mécanisme.

```
HistogramMethod(database, epsilon, a, b){
    steps = 1.5*n/log(n)

    /* composition theorem */
    epsilon /= 9;

    result = {};
    for(d in {1 ... 9}){ /* which decile */
        T = d*card(database)/10;
        for(i in {1 ... steps}){
            fi = x -> card({element in x | element < i*(b-a)/steps});
            queries.push_back(fi);
        }
        T = d*card(database)/10;
        result.push_back(AboveThreshold(database, queries, T, epsilon)
                        *(b-a)/steps});
    }
    return result;
}
```

Les entrée a et b donnent une minoration et une majoration de l'ensemble des valeurs d'entrées. L'algorithme découpe alors l'intervalle $[a, b]$ en `steps` intervalles de même tailles. Pour chaque décile, l'entier renvoyé par `Abovethreshold` est l'indice de la première valeur à dépasser ce décile.

FIGURE 1 – Le découpage pour $a = 0$, $b = 1$, `steps` $= 4$

Théorème 2.2.1.1 :
`HistogramMethod` est ε-*differentially private*.

Démonstration : Les requêtes envoyé par l'algorithme à `AboveThreshold` sont bien de sensibilité 1. Chacun des neuf appels à cette fonction est donc $\varepsilon/9$-*differentially private*. Le théorème de composition assure alors que `HistogramMethod` est ε-*differentially private*.

Maintenant que nous avons vu que cet algorithme est bien *differentially private*, nous allons essayer d'évaluer sa précision. Cela ne sera pas évident car la précision de l'algorithme dépend beaucoup du jeu de donné en entré.

Lemme 2.2.1.1 : *`AboveThreshold` est (α, β) − accurate*
Pour tout $\beta \in]0, 1[$, tout $x \in \mathcal{X}^{(\mathbb{N})}$, tout $\{f_i\}_i = Q \in \left(\mathcal{X}^{(\mathbb{N})} \to \mathcal{T}\right)^{\mathbb{N}}$, tout $\varepsilon > 0$, tout $T \in \mathbb{R}$, en posant $\alpha = 8\left(\log(k) + \log(2/\beta)\right)/\varepsilon$ et $k =$ `AboveThreshold(x, Q, T, epsilon)`, on a, en reprenant les notations de l'algorithme,

$$\mathbb{P}\left(\forall i < k \ f_i(x) + \nu_i < T + \alpha \wedge f_k(x) + \nu_k > T - \alpha\right) \geq 1 - \beta$$

Remarque : *Ce lemme est due à [**dwork2014the**]. Nous reprenons aussi la démonstration ici car la démonstration originale ne nous semble pas assez claire et trop bancale mathématiquement.*

Démonstration : Reprenons les notations de l'énoncé. Montrons déjà qu'il suffit de démontrer que

$$\mathbb{P}\left(\max_{i \leq k} |\nu_i| + |T - \hat{T}| < \alpha\right) \geq 1 - \beta \qquad (1)$$

où $\hat{T}$ est le seuil bruité défini à la ligne 4 de l'algorithme [2.1]. Or, nous avons, en posant pour tout $i \leq k$, $y_i = f_i(x)$

$$y_k + \nu_k \geq \hat{T} \overset{\text{IT}}{\geq} T - |T - \hat{T}|$$

Mutatis mutandis

$$\forall i < k \quad y_i \leq \hat{T} + |\nu_i| \leq T + |T - \hat{T}| + |\nu_i|$$

Ainsi,

$$\mathbb{P}\left(\forall i < k \;\, f_i(x) + \nu_i < T + \alpha \wedge f_k(x) + \nu_k > T - \alpha\right) \geq 1 - \beta$$

Démontrons enfin (1)! La variable aléatoire $T - \hat{T}$ suit une loi de LAPLACE de paramètre $2/\varepsilon$. Ainsi,

$$\mathbb{P}\left(|T - \hat{T}| \geq \frac{\alpha}{2} = \frac{\alpha\varepsilon}{4}\frac{2}{\varepsilon}\right) = \exp\left(-\frac{\varepsilon\alpha}{4}\right) = \exp\left(-2\left(\log k + \log\frac{2}{\beta}\right)\right) \leq \exp\left(-2\left(\log\frac{2}{\beta}\right)\right) \leq \frac{\beta}{2}$$

De même,

$$\mathbb{P}\left(\max_i |\nu_i| \geq \frac{\alpha}{2}\right) \leq \sum_{j=1}^{k} \mathbb{P}\left(|\nu_j| \geq \frac{\alpha}{2}\right) = k\exp\left(-\frac{-\alpha\varepsilon}{8}\right) = k\exp\left(-\log k - \log\frac{2}{\beta}\right) = \frac{k}{k}\frac{\beta}{2}$$

Enfin,

$$\begin{aligned}
\mathbb{P}\left(\max_{i\leq k}|\nu_i| + |T - \hat{T}| < \alpha\right) &\geq \mathbb{P}\left(\max_{i\leq k}|\nu_i| < \frac{\alpha}{2} \;\wedge\; |T - \hat{T}| < \frac{\alpha}{2}\right) \\
&= 1 - \mathbb{P}\left(\max_{i\leq k}|\nu_i| \geq \frac{\alpha}{2} \;\cup\; |T - \hat{T}| \geq \frac{\alpha}{2}\right) \\
&\geq 1 - \mathbb{P}\left(\max_{i\leq k}|\nu_i| \geq \frac{\alpha}{2}\right) - \mathbb{P}\left(|T - \hat{T}| \geq \frac{\alpha}{2}\right) \\
&\geq 1 - \frac{\beta}{2} - \frac{\beta}{2}
\end{aligned}$$

Finalement,

$$\mathbb{P}\left(\max_{i\leq k}|\nu_i| + |T - \hat{T}| < \alpha\right) \geq 1 - \beta$$

Ce qui démontre bien (1) et donc le lemme.

2.2.2 Analyse de complexité

La complexité de `AboveThreshold` est de l'ordre de la somme des complexité des requêtes sur le jeu de données d'entré. En notant n la taille de la base de donnée, les requêtes envoyé à `AboveThreshold` par `HistogramMethod` sont toute de complexité linéaire en n. Il y a au plus $\mathcal{O}(n/\log n)$ requêtes envoyées. L'algorithme a alors une complexité en $\mathcal{O}(n^2/\log n)$.

2.2.3 Analyse de précision - le cas de la distribution uniforme

Nous allons évaluer la précision de l'algorithme à l'aide de l'erreur quadratique moyenne entre la valeur renvoyé par le programme et la valeur attendue. Il y a plusieurs manière de penser ce qu'est la valeur attendue : elle pourrait être la valeur des déciles de l'échantillons d'entré. Néanmoins, elle peut tout aussi bien être l'ensemble des déciles de la loi. En effet, nous cherchons à répondre à des questions de statistique, l'entré peut-être un simple échantillon "représentatif" ; au quel cas nous sommes principalement intéressé par les réponses statistiques sur l'ensemble de la population et non juste sur notre échantillon.

Ces deux choix ont un réel sens. Nous avons d'abord essayé d'évaluer les performances de l'algorithme dans le premier cas. Les calculs était difficiles et menaient à des résultats difficilement exploitables. Nous avons donc choisi de réaliser les calculs sur la seconde option afin de pouvoir mener des calculs légèrement plus simples et ainsi avoir des résultats.

Nous allons commencer par démontrer quelques lemmes intermédiaires afin de démontrer les résultats de précision. Mais d'abord, donnons les définitions qui nous serrons utiles ici.

Définition 2.2.3.1 : *Fonction Beta incomplète*
On appel fonction beta incomplète la fonction

$$B : \begin{cases} [0,1] \times (\mathbb{R}_+^\star)^2 & \to & \mathbb{R}_+ \\ (x, \alpha, \beta) & \mapsto & \displaystyle\int_0^x t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t \end{cases}$$

Définition 2.2.3.2 : *Fonction Beta incomplète régularisée*
Pour tout $x \in [0,1]$, on appel fonction beta incomplète régularisée la fonction

$$I_x : \begin{cases} (\mathbb{R}_+^\star)^2 & \to & \mathbb{R}_+ \\ (\alpha, \beta) & \mapsto & \dfrac{B(x,\alpha,\beta)}{B(1,\alpha,\beta)} \end{cases}$$

Définition 2.2.3.3 : *Loi beta*
On appel loi beta de paramètre $(\alpha, \beta) \in \mathbb{R}_+^\star$ la loi de densité

$$f_{\alpha,\beta} : [0,1] \ni x \mapsto \frac{x^{\alpha-1}(1-x)^\beta - 1}{B(1,\alpha,\beta)}$$

Remarque : *On note directement que la fonction de répartition de la loi beta de paramètre (α, β) est la fonction $x \mapsto I_x(\alpha, \beta)$.*

Définition 2.2.3.4 : *Statistique d'ordre*
Soit X un échantillon statistique de cardinal $n \in \mathbb{N}$. Pour tout $k \in [\![1, n]\!]$ on note $X_{(i)}$ et on appel **statistique d'ordre** de rang k la k-ème plus petite valeur de l'échantillon.

Théorème 2.2.3.1 : *Loi des statistiques d'ordre d'un échantillon issue de $\mathcal{U}(0,1)$.*
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1] et $k \in [\![1, n]\!]$. La k-ème statistique d'ordre de X, $X_{(k)}$ est distribuée suivant la loi beta de paramètre $(k, n-k+1)$.

Démonstration :
TODO

Lemme 2.2.3.1 : *Estimation de l'écard entre certaines statistiques d'ordre et les déciles.*
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1] et soit $\gamma \in [0, 2d_i^l]$. Notons $(d_i^l)_i$ les déciles de la loi. Pour tout $i \in [\![1, 9]\!]$

$$\begin{aligned} \mathbb{P}\left([X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]\right) \geq & I_{d_i^l+\gamma}(in/10+\alpha, n - in/10 - \alpha + 1) \\ & - I_{d_i^l-\gamma}(in/10 - \alpha, n - in/10 + \alpha + 1) \end{aligned}$$

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1] et soit $\gamma \in [0, 2d_i^l]$. Notons $(d_i^l)_i$ les déciles de la loi. Soit enfin $i \in [\![1, 9]\!]$ et $\alpha \in]0, n/10[$. Notons que

$$\begin{aligned} \mathbb{P}\left([X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]\right) &= \mathbb{P}(X_{(in/10-\alpha)} \geq d_i^l - \gamma \ \wedge \ X_{(in/10+\alpha)} \leq d_i^l + \gamma) \\ &\geq \mathbb{P}(X_{(in/10-\alpha)} \geq d_i^l - \gamma) + \mathbb{P}(X_{(in/10+\alpha)} \leq d_i^l + \gamma) - 1 \end{aligned}$$

Or, le théorème précédent assure que

$$\begin{cases} \mathbb{P}(X_{(in/10-\alpha)} \geq d_i^l - \gamma) & = & 1 - I_{d_i^l-\gamma}(in/10 - \alpha, n - in/10 + \alpha + 1) \\ \mathbb{P}(X_{(in/10+\alpha)} \leq d_i^l + \gamma) & = & I_{d_i^l+\gamma}(in/10 + \alpha, n - in/10 - \alpha + 1) \end{cases}$$

D'où

$$\begin{aligned} \mathbb{P}\left([X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]\right) \geq & I_{d_i^l+\gamma}(in/10 + \alpha, n - in/10 - \alpha + 1) \\ & - I_{d_i^l-\gamma}(in/10 - \alpha, n - in/10 + \alpha + 1) \end{aligned}$$

La combinaison des lemmes précédents permet d'obtenir un résultat de précision utile sur `HistogramMethod`.

Théorème 2.2.3.2 : (α, β)*-précision de* `HistogramMethod` *dans le cas uniforme sur* $[0, 1]$
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Soit $\gamma \in [0, 2d_i^l]$, $i \in [\![1, 9]\!]$ et $\beta \in [0, 1]$. Notons $(d_i)_i$ les déciles de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, a, b)`, $\alpha = 8\log(3n/\beta)/\varepsilon)$ et $k = 1.5n/\log n$.

$$\begin{aligned} \mathbb{P}\left(A_i \in \left[d_i^l - \gamma - \frac{1}{k}, d_i^l + \gamma + \frac{1}{k}\right]\right) \geq & I_{d_i^l+\gamma}(in/10 + \alpha, n - in/10 - \alpha + 1) \\ & - I_{d_i^l-\gamma}(in/10 - \alpha, n - in/10 + \alpha + 1) - \beta \end{aligned}$$

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Soit $\gamma \in [0, 2d_i^l]$, $i \in [\![1, 9]\!]$ et $\beta \in [0, 1]$. Notons $(d_i)_i$ les déciles de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, a, b)`, $\alpha = 8\log(3n/\beta)/\varepsilon)$ et $k = 1.5n/\log n$.

Notons E_α l'événement "$[X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]$" Et E_{A_i} l'événement "moins de α valeurs de X séparent d_i et une valeur de X dont la distance à A_i est majorée par $1/k$". Nous avons alors

$$\mathbb{P}\left(A_i \in \left[d_i^l - \gamma - \frac{1}{k}, d_i^l + \gamma + \frac{1}{k}\right]\right) \geq \mathbb{P}\left(E_{A_i} \wedge E_\alpha\right) \geq \mathbb{P}\left(E_{A_i}\right) + \mathbb{P}\left(E_\alpha\right) - 1$$

Le lemme [1] assure que

$$\mathbb{P}(E_{A_i}) \geq 1 - \beta$$

En effet, ce lemme assure que si la réponse renvoyée ne dépassait pas le seuil, l'évaluation de la requête valait au moins $T - \alpha$ (en notant T le seuil) avec une probabilité minorée par $1 - \beta$. De plus, avec cette même probabilité, on sait que l'évaluation de l'avant dernière requête était majorée par $T + \alpha$ (toujours en notant T le seuil). Ainsi, comme $T = in/10$, E_{A_i} est de probabilité au moins $1 - \beta$.

Le lemme précédent assure alors que

$$\begin{aligned} \mathbb{P}\left(A_i \in \left[d_i^l - \gamma - \frac{1}{k}, d_i^l + \gamma + \frac{1}{k}\right]\right) \geq & I_{d_i^l+\gamma}(in/10 + \alpha, n - in/10 - \alpha + 1) \\ & - I_{d_i^l-\gamma}(in/10 - \alpha, n - in/10 + \alpha + 1) - \beta \end{aligned}$$

Ce résultat n'est pas optimal. Nous avons fait des approximations. Néanmoins, nous avons une bonne borne. Nous allons maintenant utiliser ce théorème pour obtenir un résultat très important : une majoration de l'espérance de la distance entre la valeur renvoyée par le mécanisme et un décile de la loi. Ce résultat permet de savoir quelle est l'erreur à laquelle s'attendre en pratique.

Théorème 2.2.3.3 : *Précision moyenne de* `HistogramMethod` Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Soit $\gamma \in [0, 2d_i^l]$,

$i \in [\![1,9]\!]$, $k \in \mathbb{N}$ et $\beta \in [0,1]$. Notons $(d_i)_i$ les décile de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, a, b)`, $\alpha = 8\log(3n/\beta)/\varepsilon)$ et $k = 1.5n/\log n$.

$$\begin{aligned}\mathbb{E}\left(|A_i - d_i^l|\right) \leq \frac{1}{k} + d_i^l\beta + \int_0^{d_i^l} \Big(1 - I_{d_i^l+t}(in/10+\alpha, n - in/10 - \alpha + 1)\Big)\,\mathrm{d}t \\ + \int_0^{d_i^l} I_{d_i^l-t}(in/10-\alpha, n - in/10 + \alpha + 1)\mathrm{d}t \\ + \left(1 - d_i^l - \frac{1}{k}\right)\Big(1 + \beta - I_{2d_i^l}(in/10+\alpha, n - in/10 - \alpha + 1)\Big)\end{aligned}$$

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Soit $\gamma \in [0, 2d_i^l]$, $i \in [\![1,9]\!]$, $k \in \mathbb{N}$ et $\beta \in [0,1]$. Notons $(d_i)_i$ les décile de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, a, b)`, $\alpha = 8\log(3n/\beta)/\varepsilon)$ et $k = 1.5n/\log n$. On pose enfin

$$F : \left\{\begin{array}{ccc} \mathbb{R}_+ & \to & [0,1] \\ t & \mapsto & \mathbb{P}(|A_i - d_i^l| \leq t) \end{array}\right.$$

Le théorème précédent assure que

$$\begin{aligned}\forall t \in [0, 2d_i^l] \quad F\left(t + \frac{1}{k}\right) := \mathbb{P}\left(|A_i - d_i^l| \leq t + \frac{1}{k}\right) \\ \geq I_{d_i^l+t}(in/10+\alpha, n - in/10 - \alpha + 1) \\ - I_{d_i^l-t}(in/10-\alpha, n - in/10 + \alpha + 1) - \beta\end{aligned}$$

Or, comme $F(1) = 1$,

$$\mathbb{E}\left(|A_i - d_i^l|\right) = \int_0^{\infty}(1-F(t))\mathrm{d}t = \int_0^{1/k}(1-F(t))\mathrm{d}t + \int_{1/k}^{d_i^l+1/k}(1-F(t))\mathrm{d}t + \int_{d_i^l+1/k}^{1}(1-F(t))\mathrm{d}t$$

Notons que

$$\int_0^{1/k}(1-F(t))\mathrm{d}t \leq \int_0^{1/k}\mathrm{d}t = \frac{1}{k}$$

Or, les propriétés usuelles sur les fonctions de répartition assurent que

$$\forall t \geq d_i^l \quad 1 - F(t) \leq 1 - F(d_i^l)$$

Ainsi,

$$\int_{d_i^l+1/k}^{1}(1-F(t))\mathrm{d}t \leq \left(1 - d_i^l - \frac{1}{k}\right)\Big(1 + \beta - I_{2d_i^l}(in/10+\alpha, n - in/10 - \alpha + 1)\Big)$$

Finalement, nous avons démontré que

$$\begin{aligned}\mathbb{E}\left(|A_i - d_i^l|\right) \leq \frac{1}{k} + d_i^l\beta + \int_0^{d_i^l} \Big(1 - I_{d_i^l+t}(in/10+\alpha, n - in/10 - \alpha + 1)\Big)\,\mathrm{d}t \\ + \int_0^{d_i^l} I_{d_i^l-t}(in/10-\alpha, n - in/10 + \alpha + 1)\mathrm{d}t \\ + \left(1 - d_i^l - \frac{1}{k}\right)\Big(1 + \beta - I_{2d_i^l}(in/10+\alpha, n - in/10 - \alpha + 1)\Big)\end{aligned}$$

Ce résultat est vrai pour toutes valeurs de β. Nous pourrions donc majorer notre espérance par une borne inférieure. Néanmoins cela n'aurait aucun sens ici : assez d'approximations ont étés faites pour qu'une utiliser un résultat “exacte” soit futile ; une borne inf est jolie sur le papier mais n'est en pratique que difficilement exploitable. Des calculs numériques montrent le choix $\beta = 1/(\sqrt{n}\log n)$ n'est “pas trop” éloignée de cette borne inf. Nous disposons alors du corollaire suivant.

Corollaire 2.2.3.1 : *(im)Précision moyenne de* ***`HistogramMethod`*** Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Soit $\gamma \in [0, 2d_i^l]$, $i \in [\![1, 9]\!]$ et $k \in \mathbb{N}$. Notons $(d_i)_i$ les décile de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, a, b)`, $\alpha = 8\log(3n\sqrt{n}\log n)/\varepsilon)$ et $k = 1.5n/\log n$.

$$\begin{aligned}\mathbb{E}\left(|A_i - d_i^l|\right) \leq \frac{1}{k} + \frac{d_i^l}{\sqrt{n}\log n} + \int_0^{d_i^l} \left(1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1)\right) \mathrm{d}t \\ + \int_0^{d_i^l} I_{d_i^l-t}(in/10 - \alpha, n - in/10 + \alpha + 1)\mathrm{d}t \\ + \left(1 - d_i^l - \frac{1}{k}\right)\left(1 + \frac{1}{\sqrt{n}\log n} - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1)\right)\end{aligned}$$

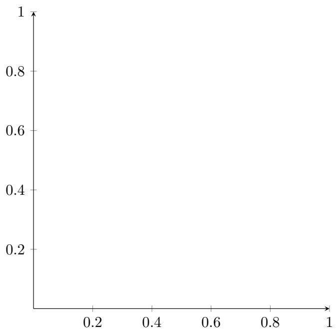


FIGURE 2 – Graphes de $\beta \mapsto \frac{1}{k} + \left(d_i^l - \frac{1}{k}\right)\beta + 2\sqrt{\frac{3\pi}{n}} + \frac{4\sqrt{\pi}}{\sqrt{n-4\alpha}}e^{-\alpha^2}$ pour $i = 5$, $n = 10000$, $k = 100$ et $\varepsilon = 1$

2.2.4 Analyse de précision - le cas de la loi normale centrée réduite

Les lois normales est très utilisées en statistique notamment car elle permettent de modéliser les phénomènes issues de plusieurs événement aléatoires. Le théorème central limite viens jouer un rôle clé dans la prépondérance de l'utilisation de ces lois. Il semble alors crucial d'étudier la précision de notre algorithme dans le cas où les données d'entré suivent une loi normale.

Le théorème de précision est très analogue à celui obtenue dans le cas uniforme. Nous ne détaillons pas ici les lemmes intermédiaires et la démonstration car il s'agit formellement de la même chose. Il est néanmoins nécessaire d'introduire quelques objets usuels en plus car la loi normale est plus complexe que la loi uniforme.

Définition 2.2.4.1 : *Fonction d'erreur*
On appel fonction d'erreur la fonction suivant :

$$\mathrm{erf} : \left\{\begin{array}{ccc} \mathbb{C} & \to & \mathbb{C} \\ z & \mapsto & \frac{2}{\sqrt{\pi}}\int_0^z \exp\left(-t^2\right)\mathrm{d}t \end{array}\right.$$

Lemme 2.2.4.1 : *Déciles de* $\mathcal{N}(0,1)$.
Les déciles de $\mathcal{N}(0,1)$, notés $(d_i^l)_i$ sont

$$\forall i \in [\![1,9]\!] \quad d_i^l = \sqrt{2}\,\mathrm{erf}^{-1}(2\times 0.1i - 1)$$

Démonstration : Soit $i \in [\![1,9]\!]$. On note que

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d_i^l} \exp\left(\frac{-t^2}{2}\right)\mathrm{d}t &= \frac{\sqrt{2}}{\sqrt{2\pi}}\int_{-\infty}^{\mathrm{erf}^{-1}(2\times 0.1i-1)} \exp\left(-t^2\right)\mathrm{d}t \\
&= \frac{1}{2}\frac{2}{\sqrt{\pi}}\int_{-\infty}^{\mathrm{erf}^{-1}(2\times 0.1i-1)} \exp\left(-t^2\right)\mathrm{d}t \\
&= \frac{1}{2}\mathrm{erf}\left(\mathrm{erf}^{-1}(2\times 0.1i-1)\right) + \frac{1}{2}\frac{2}{\sqrt{\pi}}\int_{-\infty}^{0} \exp\left(-t^2\right)\mathrm{d}t \\
&= 0.1i - \frac{1}{2} + \frac{1}{2} \\
&= 0.1i
\end{aligned}$$

La démonstration dans le cas d'une loi normale est analogue à celle du cas uniforme. Nous aurons donc des lemmes similaires. Les démonstrations seront néanmoins laissées en appendix [A].

Lemme 2.2.4.2 : *Estimation de l'écard entre les déciles empiriques et ceux de la loi normale centrée réduite.*
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite et soit $\gamma \in [0, d_i^l]$. Notons $(d_i)_i$ les déciles empiriques de X et $(d_i^l)_i$ les déciles de la loi normale centrée réduite. Pour tout $i \in [\![1,9]\!]$

$$\mathbb{P}(d_i \in [d_i^l - \gamma/2, d_i^l + \gamma/2]) \geq 1 - \eta$$

Avec

$$\eta = \exp\left(-\frac{n\gamma^2}{i^2}\left(\frac{i}{10} - \frac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-(d_i^l)^2\right)\right) + \exp\left(-\frac{5\gamma^2 in}{16\pi\left(i + 5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-(d_i^l)^2\right)\right)$$

Lemme 2.2.4.3 :
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in [\![1,9]\!]$ et $k \in \mathbb{N}$. Il y a au moins α valeurs de X dans chacun des intervalles $[d_i^l - \gamma, d_i^l - \gamma/2]$ et $[d_i^l + \gamma/2, d_i^l + \gamma]$ avec une probabilité au moins $1 - \beta$ avec

$$\beta = 2\exp\left(-\frac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right) - \frac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right)$$

Théorème 2.2.4.1 : (α,β)*-précision de* ***`HistogramMethod`*** *dans le cas de la loi normale centrée réduite*
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in [\![1,9]\!]$, $k \in \mathbb{N}$ et $\beta \in [0,1]$. Notons $(d_i)_i$ les déciles empiriques de X et $(d_i^l)_i$ les déciles de la loi normale centrée réduite. Posons A la variable aléatoire `HistogramMethod(X, epsilon, k, a, b)`.

$$\mathbb{P}\left(A_i \in [d_i^l - \gamma, d_i^l + \gamma]\right) \geq 1 - \beta - \eta - \mu$$

Avec

$$\begin{cases}
\alpha &= \dfrac{8\left(\log k + \log(2/\beta)\right)}{\varepsilon} \\
\mu &= 2\exp\left(-\dfrac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\dfrac{(|d_i^l| + \gamma)^2}{2}\right) - \dfrac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right) \\
\eta &= \exp\left(-\dfrac{n\gamma^2}{i^2}\left(\dfrac{i}{10} - \dfrac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-(d_i^l)^2\right)\right) + \exp\left(-\dfrac{5\gamma^2 in}{16\pi\left(i + 5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-(d_i^l)^2\right)\right)
\end{cases}$$

3 Le mécanisme de sensibilité inverse

3.1 Présentation du mécanisme

Le mécanisme de sensibilité inverse est introduit par Hilal Asi and John C. Duchi dans *Near Instance-Optimality in Differential Privacy* [**Asi2020NearII**]. Le mécanisme considère l'inverse du nombre de valeurs à modifié dans un ensemble de donnée pour passer à un autre ensemble de donné sur lequel la requête a une autre valeur recherchée. Cela définit alors l'utilité d'une valeur pour instancier le mécanisme exponentiel [**mcsherry2007mechanism**].

Définition 3.1.0.1 : *Longueur*
Soit $x \in \mathcal{X}^{(\mathbb{N})}$, $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $t \in \mathcal{T}$. La longueur est le nombre minimum de valeurs à modifier dans x pour obtenir x' tel que $f(x') = t$.

$$\text{len}_f(x,t) = \inf_{x' \in \mathcal{X}^{(\mathbb{N})}} \{||x - x'||_1 \mid f(x') = t\}$$

Définition 3.1.0.2 : *Mécanisme de sensibilité inverse*
Soit $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $\varepsilon \in \mathbb{R}_+$. Pour une mesure μ sur $\mathcal{T}$, on définit le mécanisme aléatoire $M(x)$ par sa fonction de densité

$$t \mapsto \frac{\exp(-\text{len}_f(x,t)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-\text{len}_f(x,s)\varepsilon/2) \mathrm{d}\mu(s)}$$

Il n'y a qu'en $f(x)$ que $\text{len}_f(x,)$ est nulle. Ainsi le dénominateur pourrait être petit est donné une grande probabilité à des valeurs distantes de $f(x)$. On [**mcsherry2007mechanism**] introduit alors une version lisse du mécanisme.

Définition 3.1.0.3 : *Longueur lisse*
Soit $x \in \mathcal{X}^{(\mathbb{N})}$, $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $\rho \in \mathbb{R}_+$. Si $\mathcal{N}$ est une norme sur $\mathcal{T}$,

$$\text{len}_f^\rho : \begin{cases} \mathcal{T} & \to & \mathbb{N} \\ t & \mapsto & \inf_{s \in \mathcal{T}, \mathcal{N}(s,t) \leq \rho} \{\text{len}_f(x,s)\} \end{cases}$$

Définition 3.1.0.4 : *Mécanisme de sensibilité inverse ρ-lisse*
Soit $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$ et $\rho, \varepsilon \in \mathbb{R}_+$. Pour une mesure μ sur $\mathcal{T}$, on définit le mécanisme aléatoire $M_{\text{cont}}(x)$ par sa fonction de densité

$$t \mapsto \frac{\exp(-\text{len}_f^\rho(x,t)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-\text{len}_f^\rho(x,s)\varepsilon/2) \mathrm{d}\mu(s)}$$

Théorème 3.1.0.1 :
Pour tout $\rho, \varepsilon \in \mathbb{R}_+$, le mécanisme de sensibilité inverse ρ-lisse est *ε-differentially private.*

Démonstration : Soit $f : \mathcal{X}^{(\mathbb{N})} \to \mathcal{T}$, $\rho, \varepsilon \in \mathbb{R}_+$, μ une mesure sur $\mathcal{T}$, $\mathcal{S} \subset \mathcal{T}$ mesurable et $x, x' \in \mathcal{X}^{(\mathbb{N})}$ voisines.

On note que

$$\begin{aligned} \mathbb{P}\left(M_{\text{cont}}(x) \in \mathcal{S}\right) &= \int_{\mathcal{S}} \frac{\exp(-\text{len}_f^\rho(x,t)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-\text{len}_f^\rho(x,s)\varepsilon/2) \mathrm{d}\mu(s)} \mathrm{d}\mu(t) \\ &\leq \int_{\mathcal{S}} \frac{\exp(-(\text{len}_f^\rho(x',t) - 1)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-(\text{len}_f^\rho(x',s) + 1)\varepsilon/2) \mathrm{d}\mu(s)} \mathrm{d}\mu(t) \\ &= \frac{\exp(\varepsilon/2)}{\exp(-\varepsilon/2)} \int_{\mathcal{S}} \frac{\exp(-\text{len}_f^\rho(x',t)\varepsilon/2)}{\int_{\mathcal{T}} \exp(-\text{len}_f^\rho(x,s)\varepsilon/2) \mathrm{d}\mu(s)} \mathrm{d}\mu(t) \\ &= \exp(\varepsilon)\mathbb{P}\left(M_{\text{cont}}(x') \in \mathcal{S}\right) \end{aligned}$$

3.2 Précision du mécanisme de sensibilité inverse pour l'estimation de déciles

L'article présentant le mécanisme de sensibilité inverse [**Asi2020NearII**] détail une borne de précision sur la médiane. Nous allons ici étendre cette démonstration au cas des déciles. Dans cette section nous nous plaçons dans le cas où les données sont identiquement distribuées à partir d'une loi ayant une distribution continue π_P au voisinage de ses déciles $(d_i^l)_i$. Commençons par énoncé le résultat.

Théorème 3.2.0.1 :
Soit $\gamma \in \mathbb{R}_+^\star$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ et $X \in [0, R]^n$ dont les éléments sont obtenues à partir d'une loi P de densité π_P continue au voisinage de ses déciles. On pose $p_{\min,i} = \inf_{t \in [d_i^l - 2\gamma, d_i^l + 2\gamma]} \pi_P(t)$. On note $(d_i)_i$ les déciles empirique de X et $(d_i^l)_i$ les déciles de la loi. Notons alors enfin M_{cont} le mécanisme de sensibilité inverse ρ-lisse.

$$\mathbb{P}\left(|M_{\text{cont},i} - d_i| > 2u + \rho\right) \leq \frac{R}{2\rho} \exp\left(-\frac{np_{\min,i} u\varepsilon}{4}\right) + 4\exp\left(-\frac{n\gamma^2 p_{\min,i}^2}{8}\right) + \frac{2\gamma}{u} \exp\left(-\frac{np_{\min,i} u}{8}\right)$$

démonstration : Ce théorème donne une borne exponentielle sur la précision de l'algorithme. Le démonstration est longue.

Découpons l'intervalle $[d_i^l - \gamma, d_i^l + \gamma]$ en intervalles $(I_j)_j$ de taille u. Pour tout j, on pose $N_j = \# I_j$. On note alors A l'événement "pour tout j, $N_j \geq nup_{\min,i}/2$" et B_i l'événement "$|d_i^l - d_i| \geq \gamma/2$".

$$\begin{aligned}
\mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho\right) &= \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A \wedge B_i\right) \mathbb{P}\left(A \wedge B_i\right) \\
&\quad + \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid \overline{A} \vee \overline{B}_i\right) \mathbb{P}\left(\overline{A} \vee \overline{B}_i\right) \\
&\leq \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A \wedge B_i\right) + \mathbb{P}\left(\overline{A} \vee \overline{B}_i\right) \\
&= \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A \wedge B_i\right) + \mathbb{P}\left((\overline{A} \wedge B_i) \vee \overline{B}_i\right) \\
&\leq \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A \wedge B_i\right) + \mathbb{P}\left(\overline{A} \wedge B_i\right) + \mathbb{P}\left(\overline{B}_i\right) \\
&= \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A \wedge B_i\right) + \mathbb{P}\left(\overline{A} \mid B_i\right) \mathbb{P}(B_i) + \mathbb{P}\left(\overline{B}_i\right) \\
&\leq \mathbb{P}\left(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A \wedge B_i\right) + \mathbb{P}\left(\overline{A} \mid B_i\right) + \mathbb{P}\left(\overline{B}_i\right)
\end{aligned}$$

Nous savons que si les événements A et B surviennent, pour tout t tel que $|t - d_i| > 2u$, au moins $nup_{\min,i}/2$ éléments séparent d_i et t. Pour de tels t nous avons alors $\text{len}_f(x, t) \geq nup_{\min,i}/2$. Ainsi, pour tout s tel que $|s - d_i| > 2u + \rho$, $\text{len}_f^\rho(x, s) \geq nup_{\min,i}/2$. Enfin, pour tout t tel que $|t - d_i| > 2u + \rho$,

$$\begin{aligned}
\pi_P\left(t \mid A \wedge B\right) &= \frac{\exp\left(-\text{len}_f^\rho(x, t)\varepsilon/2\right)}{\int_{\mathcal{T}} \exp\left(-\text{len}_f^\rho(x, s)\varepsilon/2\right) \mathrm{d}\mu(s)} \\
&\leq \frac{\exp\left(-nup_{\min,i}\varepsilon/4\right)}{\int_{\mathcal{T}} \exp\left(-\text{len}_f^\rho(x, s)\varepsilon/2\right) \mathrm{d}\mu(s)} \\
&\leq \frac{\exp\left(-nup_{\min,i}\varepsilon/4\right)}{\int_{d_i-\rho}^{d_i+\rho} \exp\left(-\text{len}_f^\rho(x, s)\varepsilon/2\right) \mathrm{d}\mu(s)} \\
&= \frac{\exp\left(-nup_{\min,i}\varepsilon/4\right)}{\int_{d_i-\rho}^{d_i+\rho} \mathrm{d}\mu(s)} \\
&= \frac{\exp\left(-nup_{\min,i}\varepsilon/4\right)}{2\rho}
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\mathbb{P}\left(|M_{\text{cont}} - d_i| > 2u + \rho \mid A \wedge B_i\right) &\leq \int_{\mathcal{T}} \frac{\exp\left(-nup_{\min,i}\varepsilon/4\right)}{2\rho} \mathbb{1}_{|t-d_i|>2u+\rho} \mathrm{d}\mu(t) \\
&\leq \frac{\exp\left(-nup_{\min,i}\varepsilon/4\right)}{2\rho} \mu(\mathcal{T}) \\
&= \frac{R}{2\rho} \exp\left(-nup_{\min,i}\varepsilon/4\right)
\end{aligned}$$

Nous allons maintenant calculer la probabilité de l'événement $\overline{B}_i$. Pour cela, on pose $\alpha = \gamma/2$, pour tout $j \in [\![0, n-1]\!]$ on pose $C_j^i = \mathbb{1}_{x_i > d_i^l + \alpha}$ et $C^i = \sum_{j=0}^{n-1} C_j$. L'événement C^i dénote le nombre d'éléments de X plus grands que $d_i^l + \alpha$. Par définition de $p_{\min,i}$ assure que

$$\begin{aligned}
\hat{p} \quad &:= \quad \mathbb{P}(C_j^i = 1) \\
&= \quad 1 - \int_0^{d_i^l} \pi_P(t)\mathrm{d}\mu(t) - \int_{d_i^l}^{d_i^l+\alpha} \pi_P(t)\mathrm{d}\mu(t) \\
&\overset{\text{déf de } d_i^l}{=} \quad 1 - \frac{i}{10} - \int_{d_i^l}^{d_i^l+\alpha} \pi_P(t)\mathrm{d}\mu(t) \\
&\leq \quad \frac{10-i}{10} - p_{\min,i} \int_{d_i}^{d_i^l+\alpha} \mathrm{d}\mu(t) \\
&= \quad \frac{10-i}{10} - \alpha p_{\min,i}
\end{aligned}$$

Or, si $d_i > d_i^l$, $C^i \geq in/10$. Ainsi, en utilisant une borne de Chernoff (C^i est d'espérance $\hat{p}n$ et les $(C_j^i)_j$ sont indépendantes),

$$\begin{aligned}
\mathbb{P}\left(d_i > d_i^l + \alpha\right) &\leq \mathbb{P}\left(C^i \geq \frac{in}{10}\right) \\
&= \mathbb{P}\left(\sum_{j=0}^{n-1} C_j^i \geq \hat{p}n\left(1 - \left(1 - \frac{i}{\hat{p}10}\right)\right)\right) \\
&\leq \exp\left(-\left(1 - \frac{i}{\hat{p}10}\right)^2 \frac{n\hat{p}}{2}\right) \\
&= \exp\left(-\left(\hat{p} - \frac{i}{10}\right)^2 \frac{n}{2\hat{p}}\right) \\
&\leq \exp\left(-\left(\alpha p_{\min,i}\right)^2 \frac{n}{2\hat{p}}\right) \\
&\leq \exp\left(-\alpha^2 p_{\min,i}^2 \frac{n}{i/5 - 2\alpha p_{\min,i}}\right) \\
&\leq \exp\left(-\frac{1}{2}\alpha^2 p_{\min,i}^2 n\right)
\end{aligned}$$

On montre alors de même que $\mathbb{P}\left(d_i < d_i^l - \alpha\right) < \exp\left(-\frac{1}{2}\alpha^2 p_{\min,i}^2 n\right)$. Nous avons donc montré que

$$\mathbb{P}\left(B_i\right) \geq 1 - 2\exp\left(-\frac{1}{8} n\gamma^2 p_{\min,i}^2\right)$$

Finalement, il ne nous reste plus qu'à minorer $\mathbb{P}(A \mid B_i)$! Pour cela, notons que

$$\mathbb{P}(A \mid B_i) \geq (A \mid B_i)\mathbb{P}(B_i) = \mathbb{P}(A) - \mathbb{P}\left(A \wedge \overline{B}_i\right) \geq \mathbb{P}(A) - \mathbb{P}(\overline{B}_i)$$

Pour tout $k \leq n-1$ on pose alors $Z_k = \mathbb{1}_{x_k \in I_j}$ et on a $N_j = \sum_{k=0}^{n-1} Z_k$. On note que $\mathbb{P}(Z_j = 1) \geq up_{\min,i}$. Utiliser une nouvelle fois une borne de CHERNOFF assure enfin que

$$\mathbb{P}\left(N_j < nup_{\min,i}/2\right) = \mathbb{P}\left(N_j < nup_{\min,i}\left(1 - \frac{1}{2}\right)\right) < \exp\left(-\frac{1}{8}nup_{\min,i}\right)$$

Enfin,

$$\mathbb{P}\left(\overline{A}\right) = \mathbb{P}\left(\bigcup_{j=0}^{2\gamma/u} N_j < nup_{\min,i}/2\right) \leq \sum_{j=0}^{2\gamma/u} \mathbb{P}\left(N_j < nup_{\min,i}/2\right) \leq \frac{2\gamma}{u}\exp\left(-\frac{1}{8}nup_{\min,i}\right)$$

On obtient alors

$$\mathbb{P}(A \mid B_i) \geq 1 - \frac{2\gamma}{u}\exp\left(-\frac{1}{8}nup_{\min,i}\right) - 2\exp\left(-\frac{1}{8}n\gamma^2 p_{\min,i}^2\right)$$

Ce que nous permet alors d'obtenir le résultat recherché !

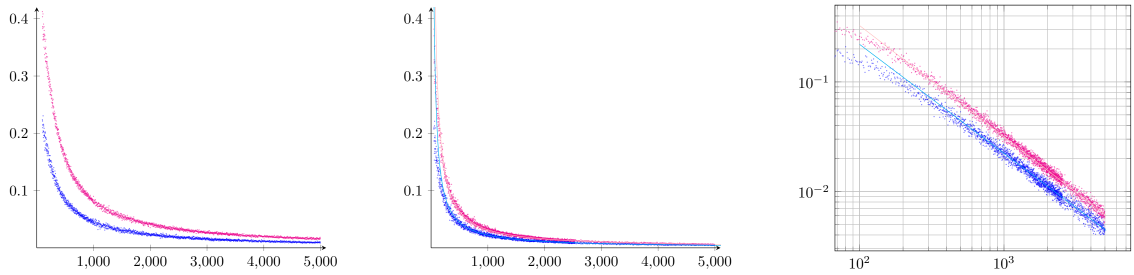
4 Comparaison entre le mécanisme de sensibilité inverse et la méthode des histogrammes

Dans les sections précédentes nous avons présenté la méthode de sensibilité inverse ainsi que la méthode que nous avons introduite, la méthode des histogrammes. Nous avons étudié en détail notre méthode et nous avons reporté une partie de l'étude du mécanisme de sensibilité inverse et nous avons produit des résultats supplémentaires. Ces deux méthodes présentent de bonnes bornes de précisions tout en étant ε-differentially private.

Pour cette comparaison nous avons décidé de nous concentrer sur deux aspects principaux aspects : la précisions des algorithmes pour des lois usuelles et l'influence du choix de ε sur la précision avec des données réelles. Les lois usuelles étudiées sont la loi uniforme sur $[0,1]$ et la loi normale centrée réduite. Nous avons choisis ces deux lois car elles modélisent de nombreux phénomènes courants et que les lois normales ont une importance particulière en statistique grâce au théorème central limite.

4.1 Résultat dans le cas de la loi uniforme standard

Nous avons calculé l'écart quadratique moyen en fonction de la taille de l'échantillon dans le cas de la loi uniforme standard. Pour cela, pour tout $n \in [\![100, 5000]\!]$ nous avons lancé les deux algorithmes sur 50 ensembles de données indépendants et identiquement distribué suivant $\mathcal{U}(0,1)$.



(a) $\varepsilon = 0.5$ (b) $\varepsilon = 1$ (c) $\log - \log$: $\varepsilon = 1$

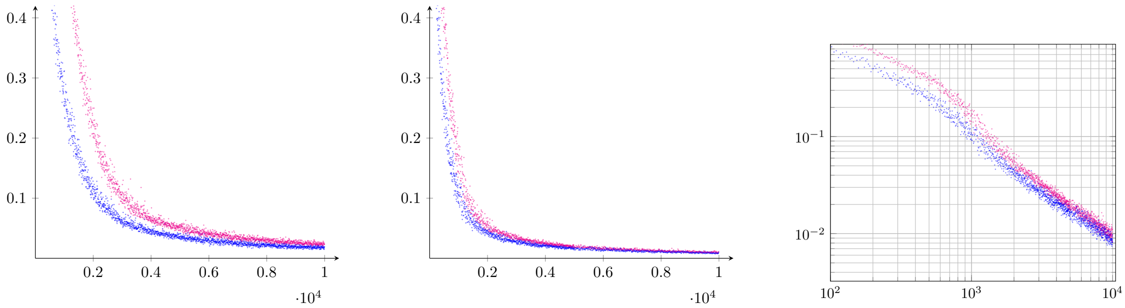
FIGURE 3 – Écart-quadratique moyen sur le calcul des déciles en fonction de n (la taille de l'échantillon). La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

Le graphe $\log - \log$ montre que dans le cas $\varepsilon = 1$, l'écart quadratique semble être d'espérance $35n^{-1.015}$ pour la méthode des histogrammes et $21.5n^{-0.995}$ pour le mécanisme de sensibilité inverse. On observe alors que pour des valeurs de n courantes ($\leq 10^8$), le mécanisme de sensibilité inverse semble meilleur que la méthode que nous avons introduite et que **notre méthode est asymptotiquement meilleure** même si cela ne sera pas le cas en pratique.

Enfin, les deux mécanismes offrent vraiment des performances similaires. Le mécanisme de sensibilité inverse devrait être privilégié pour obtenir une meilleure précision. Néanmoins, la méthode des histogrammes est une alternative viable.

4.2 Résultat dans le cas de la loi normale centrée réduite

Nous avons suivi la même méthodologie que dans le cas de la loi uniforme standard. Les résultats obtenues sont similaires. les résultats suivent moins uniformément le modèle d'une puissance mais semble aussi y coller. Comme dans le cas uniforme, l'écart quadratique est moins bon avec la méthode des histogrammes. Néanmoins, cet écart semble asymptotiquement meilleur par rapport à celui obtenue avec le mécanisme de sensibilité inverse.



(a) $\varepsilon = 0.5$ (b) $\varepsilon = 1$ (c) $\log - \log : \varepsilon = 1$

FIGURE 4 – Écart-quadratique moyen sur le calcul des déciles en fonction de n (la taille de l'échantillon). La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

4.3 Résultat sur des données réelles

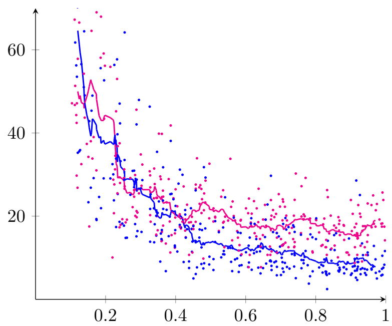


FIGURE 5 – Écart-quadratique sur le calcul des déciles en fonction de ε. La méthode des histogrammes est en magenta et le mécanisme de sensibilité inverse est en bleu.

Nous avons décidé de comparer les résultats des deux méthodes sur une base de donnée réelle [**salaries**] afin de voir comment les méthodes se comportent vraiment en pratique. Pour cela nous avons utilisé les salaires annuels des agents du publique de l'état de New-York (États-Unis). Ces données ont directement été publiques par une agence gouvernementale, ce qui assure une certaine fiabilité. Nous travaillons alors sur un ensemble de près de 400 000 salaires annuels.

Les courbes sur le graphe 5 sont des SMA_{10} (*simple moving average* de paramètre 10). Cet indicateur permet de lisser les fluctuations locales afin de mettre en avant les tendances globales. Ainsi, le graphe 5 montre que la méthode de sensibilité inverse est globalement plus précise. Il n'y a toute fois pas d'ordre de grandeur de différence entre les erreurs de deux algorithmes. Leurs performances sont donc similaires.

De plus, ces deux algorithmes fournissent des résultats précis. En effet, les déciles du jeu de donné sont 34 902, 38 574, 41 848, 46 862, 56 844, 67 121, 75 254, 84 751 et 99 637. L'erreur quadratique observée, proche de 20 est donc négligeable au vu des ordres de grandeurs des données.

A HistogramMethod : Analyse de précision - le cas de la loi normale centrée réduite

A.1 Démonstration du lemme [??]

Lemme A.1.0.1 : *Estimation de l'écard entre les déciles empiriques et ceux de la loi normale centrée réduite.*
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite et soit $\gamma \in [0, d_i^l]$. Notons $(d_i)_i$ les déciles empiriques de X et $(d_i^l)_i$ les déciles de la loi normale centrée réduite. Pour tout $i \in [\![1, 9]\!]$

$$\mathbb{P}(d_i \in [d_i^l - \gamma/2, d_i^l + \gamma/2]) \geq 1 - \eta$$

Avec

$$\eta = \exp\left(-\frac{n\gamma^2}{i^2}\left(\frac{i}{10} - \frac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-(d_i^l)^2\right)\right) + \exp\left(-\frac{5\gamma^2 in}{16\pi\left(i + 5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-(d_i^l)^2\right)\right)$$

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur [0,1]. Notons $(d_i)_i$ les déciles empiriques de X et $(d_i^l)_i$ ceux de la loi. Soit $\gamma \in [0, d_i^l]$. On note que

$$\begin{aligned}\mathbb{P}(d_i \in [d_i^l - \gamma/2, d_i^l + \gamma + 2]) &= 1 - \mathbb{P}(d_i \notin [d_i^l - \gamma/2, d_i^l + \gamma/2]) \\ &= 1 - \mathbb{P}(d_i \leq d_i^l - \gamma/2 \vee d_i \geq d_i^l + \gamma/2)\end{aligned}$$

On pose alors A = "il y a au moins $in/10$ valeurs plus petites que $d_i^l - \gamma/2$" et B = "il y a au plus $in/10$ valeurs plus petites que $d_i^l + \gamma/2$". Pour tout $j \in [\![0, n-1]\!]$ on pose $A_j = \mathbb{1}_{x_j \leq d_i^l + \gamma/2}$, $B_j = \mathbb{1}_{x_j \leq d_i^l + \gamma/2}$, $A_s = \sum_{j=0}^{n-1} A_j$ et $B_s = \sum_{j=0}^{n-1} B_j$. On a alors, $A = \{A_s \geq in/10\}$ et $B = \{B_s \leq in/10\}$. Une application d'une borne de CHERNOFF assure alors que

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_s \geq in/10) \\ &= \mathbb{P}\left(A_s \geq \frac{n}{\sqrt{2\pi}}\int_{-\infty}^{d_i^l - \gamma/2}\exp\left(\frac{-t^2}{2}\right)\mathrm{d}t\left(1 + \frac{i\sqrt{2\pi}}{10\int_{-\infty}^{d_i^l-\gamma/2}\exp(-t^2/2)\,\mathrm{d}t} - 1\right)\right) \\ &\overset{d_i^l \geq \gamma}{\leq} \exp\left(-\frac{n}{3\sqrt{2\pi}}\int_{-\infty}^{d_i^l-\gamma/2}\exp\left(\frac{-t^2}{2}\right)\mathrm{d}t\left(\frac{i\sqrt{2\pi}}{10\int_{-\infty}^{d_i^l-\gamma/2}\exp(-t^2/2)\,\mathrm{d}t} - 1\right)^2\right) \\ &= \exp\left(-\frac{n}{3}\left(\frac{i}{10} - \frac{1}{\sqrt{2\pi}}\int_{d_i^l-\gamma/2}^{d_i^l}\exp\left(\frac{-t^2}{2}\right)\mathrm{d}t\right)\left(\frac{i\sqrt{2\pi}}{10\int_{-\infty}^{d_i^l-\gamma/2}\exp(-t^2/2)\,\mathrm{d}t} - 1\right)^2\right) \\ &\leq \exp\left(-\frac{n}{3}\left(\frac{i}{10} - \frac{\gamma}{2\sqrt{2\pi}}\right)\left(\frac{i\sqrt{2\pi}}{10\int_{-\infty}^{d_i^l-\gamma/2}\exp(-t^2/2)\,\mathrm{d}t} - 1\right)^2\right)\end{aligned}$$

Or, la valeurs des déciles de la loi normale centrée réduite étant connues 1,

$$
\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d_i^l-\gamma/2}\exp\left(\frac{-t^2}{2}\right)\mathrm{d}t &= \frac{1}{2}\frac{2}{\sqrt{\pi}}\int_{-\infty}^{(d_i^l-\gamma/2)/\sqrt{2}}\exp\left(-t^2\right)\mathrm{d}t\\
&= \frac{1}{2}+\frac{1}{2}\operatorname{erf}\left(\frac{d_i^l-\gamma/2}{\sqrt{2}}\right)\\
&= \frac{1}{2}+\frac{1}{2}\operatorname{erf}\left(\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)-\frac{\gamma}{2\sqrt{2}}\right)\\
&= \frac{1}{2}+\frac{1}{2}\operatorname{erf}\left(\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)\right)-\frac{1}{\sqrt{\pi}}\int_{\operatorname{erf}^{-1}(2\times 0.1i-1)-\gamma/2\sqrt{2}}^{\operatorname{erf}^{-1}(2\times 0.1i-1)}\exp\left(-t^2\right)\mathrm{d}t\\
&= \frac{i}{10}-\frac{1}{\sqrt{\pi}}\int_{\operatorname{erf}^{-1}(2\times 0.1i-1)-\gamma/2\sqrt{2}}^{\operatorname{erf}^{-1}(2\times 0.1i-1)}\exp\left(-t^2\right)\mathrm{d}t\\
&\leq \frac{i}{10}-\frac{\gamma}{2\sqrt{2\pi}}\exp\left(-\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)^2\right)
\end{aligned}
$$

Enfin, comme $25/(6\pi)\geq 5$,

$$
\begin{aligned}
\mathbb{P}(A) &\leq \exp\left(-\frac{n}{3}\left(\frac{i}{10}-\frac{\gamma}{2\sqrt{2\pi}}\right)\left(\frac{i}{i-5\gamma/\sqrt{2\pi}\exp\left(-\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)^2\right)}-1\right)^2\right)\\
&\leq \exp\left(-\frac{n\gamma^2}{i^2}\left(\frac{i}{10}-\frac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-2\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)^2\right)\right)\\
&= \exp\left(-\frac{n\gamma^2}{i^2}\left(\frac{i}{10}-\frac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-(d_i^l)^2\right)\right)
\end{aligned}
$$

Finalement,

$$
\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}\left(B_s\leq in/10\right)\\
&= \mathbb{P}\left(B_s\leq\frac{n}{\sqrt{2\pi}}\left(\int_{-\infty}^{d_i^l+\gamma/2}\exp\left(-\frac{t^2}{2}\mathrm{d}t\right)\right)\left(1-\left(1-\frac{i\sqrt{2\pi}}{10\int_{-\infty}^{d_i^l+\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t}\right)\right)\right)\\
&\leq \exp\left(-\frac{n}{2\sqrt{2\pi}}\left(\int_{-\infty}^{d_i^l+\gamma/2}\exp\left(-\frac{t^2}{2}\mathrm{d}t\right)\right)\left(1-\frac{i\sqrt{2\pi}}{10\int_{-\infty}^{d_i^l+\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t}\right)^2\right)\\
&\leq \exp\left(-\frac{in}{20}\left(1-\frac{i}{i+5\gamma/\sqrt{2\pi}\exp\left(-(\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)+\gamma/2)^2\right)}\right)^2\right)\\
&= \exp\left(-\frac{25\gamma^2 in}{40\pi}\left(\frac{\exp\left(-(\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)+\gamma/2)^2\right)}{i+5\gamma/\sqrt{2\pi}\exp\left(-(d_i^l+\gamma/2)^2\right)}\right)^2\right)\\
&\leq \exp\left(-\frac{5\gamma^2 in}{16\pi\left(i+5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-2(\operatorname{erf}^{-1}\left(2\times 0.1i-1\right)+\gamma/2)^2\right)\right)\\
&\leq \exp\left(-\frac{5\gamma^2 in}{16\pi\left(i+5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-(d_i^l)^2\right)\right)
\end{aligned}
$$

A.2 Démonstration du lemme [3]

Lemme A.2.0.1 :

Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in [\![1, 9]\!]$ et $k \in \mathbb{N}$. Il y a au moins α valeurs de X dans chacun des intervalles $[d_i^l - \gamma, d_i^l - \gamma/2]$ et $[d_i^l + \gamma/2, d_i^l + \gamma]$ avec une probabilité au moins $1 - \beta$ avec

$$\beta = 2\exp\left(-\frac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right) - \frac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right)$$

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in [\![1, 9]\!]$ et $\alpha \in \mathbb{N}$. On pose $A =$ “il y a au moins α valeurs dans l'intervalle $[d_i^l - \gamma, d_i^l - \gamma/2]$” et $B =$ “il y a au moins α valeurs dans l'intervalle $[d_i^l + \gamma/2, d_i^l + \gamma]$”. Pour tout $j \in [\![0, n-1]\!]$ on pose $A_j = \mathbb{1}_{x_j \in [d_i^l - \gamma, d_i^l - \gamma/2]}$, $B_j = \mathbb{1}_{x_j \in [d_i^l + \gamma/2, d_i^l + \gamma]}$, $A_s = \sum_{j=0}^{n-1} A_j$ et $B_s = \sum_{j=0}^{n-1} B_j$. On a alors, $A = \{A_s \geq \alpha\}$ et $B = \{B_s \geq \alpha\}$.

$$\begin{aligned}
\mathbb{P}(A \ \wedge\ B) &= \mathbb{P}(A_s \geq \alpha \ \wedge\ B_s \geq \alpha) \\
&\geq \mathbb{P}(A_s \geq \alpha) + \mathbb{P}(B_s \geq \alpha) - 1 \\
&= 1 - \mathbb{P}(A_s < \alpha) - \mathbb{P}(B_s < \alpha)
\end{aligned}$$

Une application d'une borne de Chernoff assure alors que

$$\begin{aligned}
\mathbb{P}(A_s < \alpha) &= \mathbb{P}\left(A_s < \frac{n}{\sqrt{2\pi}}\int_{d_i^l - \gamma}^{d_i^l - \gamma/2} \exp\left(-\frac{t^2}{2}\right)\mathrm{d}t\left(1 - \left(1 - \frac{\alpha\sqrt{2\pi}}{n\int_{d_i^l-\gamma}^{d_i^l-\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t}\right)\right)\right) \\
&\leq \exp\left(-\frac{n}{2\sqrt{2\pi}}\int_{d_i^l-\gamma}^{d_i^l-\gamma/2}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t\left(1 - \frac{\alpha\sqrt{2\pi}}{n\int_{d_i^l-\gamma}^{d_i^l-\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t}\right)^2\right) \\
&\leq \exp\left(-\frac{n\gamma}{4\sqrt{2\pi}}\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right)\left(\frac{n\int_{d_i^l-\gamma}^{d_i^l-\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t - \alpha\sqrt{2\pi}}{n\int_{d_i^l-\gamma}^{d_i^l-\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t}\right)^2\right) \\
&\leq \exp\left(-\frac{1}{n\gamma\sqrt{2\pi}}\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right)\left(n\int_{d_i^l-\gamma}^{d_i^l-\gamma/2}\exp\left(-t^2/2\right)\mathrm{d}t - \alpha\sqrt{2\pi}\right)^2\right) \\
&\leq \exp\left(-\frac{n}{\gamma\sqrt{2\pi}}\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right)\left(\frac{\gamma}{2}\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right) - \frac{\alpha\sqrt{2\pi}}{n}\right)^2\right) \\
&\leq \exp\left(-\frac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right) - \frac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right)
\end{aligned}$$

Nous pourrions alors montrer, exactement de la même manière que

$$\mathbb{P}(B_s < \alpha) \leq \exp\left(-\frac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right) - \frac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right)$$

Finalement,

$$\mathbb{P}(A \ \wedge\ B) \geq 1 - 2\exp\left(-\frac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\frac{(|d_i^l| + \gamma)^2}{2}\right) - \frac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right)$$

A.3 Démonstration du théorème [1]

Théorème A.3.0.1 : (α, β)*-précision de* `HistogramMethod` *dans le cas de la loi normale centrée réduite*
Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in [\![1, 9]\!]$, $k \in \mathbb{N}$ et $\beta \in [0, 1]$. Notons $(d_i)_i$ les déciles empiriques de X et $(d_i^l)_i$ les déciles de la loi normale centrée réduite. Posons A la variable aléatoire `HistogramMethod(X, epsilon, k, a, b)`.

$$\mathbb{P}\left(A_i \in [d_i^l - \gamma, d_i^l + \gamma]\right) \geq 1 - \beta - \eta - \mu$$

Avec

$$\begin{cases} \alpha & = \dfrac{8\left(\log k + \log(2/\beta)\right)}{\varepsilon} \\ \mu & = 2\exp\left(-\dfrac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\dfrac{(|d_i^l| + \gamma)^2}{2}\right) - \dfrac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right) \\ \eta & = \exp\left(-\dfrac{n\gamma^2}{i^2}\left(\dfrac{i}{10} - \dfrac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-(d_i^l)^2\right)\right) + \exp\left(-\dfrac{5\gamma^2 in}{16\pi\left(i + 5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-(d_i^l)^2\right)\right) \end{cases}$$

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in [\![1, 9]\!]$, $k \in \mathbb{N}$ et $\beta \in [0, 1]$. Notons $(d_i)_i$ les déciles empiriques de X et $(d_i^l)_i$ les déciles de la loi normale centrée réduite. Posons A la variable aléatoire `HistogramMethod(X, epsilon, k, a, b)`.

On pose

$$\alpha = \frac{8\left(\log k + \log(2/\beta)\right)}{\varepsilon}$$

Notons alors E_α l'événement "Il y a au moins α valeurs de X dans chacun des intervalles $[d_i^l - \gamma, d_i^l - \gamma/2]$ et $[d_i^l + \gamma/2, d_i^l + \gamma]$" Et E_{A_i} l'événement "moins de α valeurs de X séparent d_i et A_i". Nous avons alors

$$\begin{aligned} \mathbb{P}\left(A_i \in [d_i^l - \gamma, d_i^l + \gamma]\right) &\geq \mathbb{P}\left(E_{A_i} \wedge E_\alpha \wedge d_i \in [d_i^l - \gamma/2, d_i^l + \gamma/2]\right) \\ &\geq \mathbb{P}\left(E_{A_i}\right) + \mathbb{P}\left(E_\alpha\right) + \mathbb{P}\left(d_i \in [d_i^l - \gamma/2, d_i^l + \gamma/2]\right) - 2 \end{aligned}$$

Les lemmes précédent assurent alors que

$$\begin{aligned} \mathbb{P}\left(A_i \in [0.1i - \gamma, 0.1i + \gamma]\right) &\geq (1 - \beta) + (1 - \mu) + (1 - \eta) - 2 \\ &\geq 1 - \beta - \mu - \eta \end{aligned}$$

Avec

$$\begin{cases} \alpha & = \dfrac{8\left(\log k + \log(2/\beta)\right)}{\varepsilon} \\ \mu & = 2\exp\left(-\dfrac{n\gamma}{4\sqrt{2\pi}}\left(\exp\left(-\dfrac{(|d_i^l| + \gamma)^2}{2}\right) - \dfrac{2\alpha\sqrt{2\pi}}{n\gamma}\right)^3\right) \\ \eta & = \exp\left(-\dfrac{n\gamma^2}{i^2}\left(\dfrac{i}{10} - \dfrac{\gamma}{2\sqrt{2\pi}}\right)\exp\left(-(d_i^l)^2\right)\right) + \exp\left(-\dfrac{5\gamma^2 in}{16\pi\left(i + 5\gamma/\sqrt{2\pi}\right)^2}\exp\left(-(d_i^l)^2\right)\right) \end{cases}$$