

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 Supérieure de Lyon
France
25 juillet 2022

Table des matières

Table des matières	1
1 Introduction	2
1.1 Présentation du problème	2
1.1.1 De l'importance de respecter la confidentialité	2
1.1.2 Anonymiser les données n'est pas suffisant	2
1.1.3 L'appel à la <i>differential privacy</i>	3
1.1.4 Contenu de ce rapport	3
1.2 Background essentiel sur la <i>differential privacy</i>	3
1.3 L'échec de la méthode naïve	5
2 Méthode des histogrammes	5
2.1 AboveThreshold	5
2.2 Présentation de la méthode des histogrammes	7
2.3 Analyse de complexité	9
2.4 Analyse de précision - le cas de la distribution uniforme standard	9
2.4.1 Analyse de précision : borne exacte	9
2.4.2 Analyse de précision : borne asymptotique	12
2.4.3 Analyse de précision : résultats expérimentaux	15
2.5 Analyse de précision - le cas de la loi normale centrée réduite	15
3 Le mécanisme de sensibilité inverse	16
3.1 Présentation du mécanisme	16
3.2 Précision du mécanisme de sensibilité inverse pour l'estimation de déciles	17
3.2.1 Le cas général	17
3.3 Analyse de précision - le cas de la loi uniforme standard	19
4 Comparaison entre le mécanisme de sensibilité inverse et la méthode des histogrammes	20
4.1 Comparaison des bornes obtenues	20
4.2 Résultats expérimentaux	20
4.2.1 Le cas de la loi uniforme standard	20
4.2.2 La loi normale centrée réduite	21
4.2.3 Des données réelles	21
Références	i
A Démonstration de théorèmes utiles	i
A.1 Loi des statistiques d'ordre	i
A.2 Inégalité d'HOEFFDING	i
A.3 Bornes multiplicatives de sc Chernoff	i
A.4 Déciles de la loi normale centrée réduite	i
B Le mécanisme de sensibilité inverse - probabilité de précision sur les déciles empiriques	i
C HistogramMethod : Analyse de précision - le cas de la loi normale centrée réduite	iii

1 Introduction

L'estimation de quantiles a de nombreux intérêts. Ils interviennent notamment en machine learning grâce à la régression de quantiles, ils permettent d'approximer des lois, sont utilisés dans de nombreux algorithmes tels que l'extrem gradient boosting. Durant ce stage nous avons donc décidé de nous concentrer sur l'approximation de quantiles et en particulier l'estimation de déciles.

*Nous avons proposé une méthode que nous appelons **méthode des histogrammes**. Cette méthode permet d'estimer les quantiles d'un jeu de données de manière differentially private tout en assurant un niveau de précisions de qualité. Nous avons étudié la précision de cet algorithme de manière théorique et expérimentale. Le meilleur algorithme connue à ce jour est le **mécanisme de sensibilité inverse** [AD20]. Nous avons donc aussi étudié en parti ce mécanisme et fournis des bornes qui n'était pas étudiées au par avant. Enfin, nous avons comparé ces deux mécanismes d'un point de vue théorique et expérimental.*

1.1 Présentation du problème

1.1.1 De l'importance de respecter la confidentialité

Le respect de la confidentialité est un problème majeur à l'air d'internet. *Forbes* écrivait en 2019 que la confidentialité des données sera la plus grande problématique de la prochaine décennie [Mee19]. Nous pouvons retrouver une peur au sein de la population concernant la gestion des données. En effet, selon *Pew Research Center* : “79% of adults assert they are very or somewhat concerned about how companies are using the data they collect about them, while 64% say they have the same level of concern about government data collection” et “a majority think the potential risks of data collection outweigh the benefits” [Aux+19].

1.1.2 Anonymiser les données n'est pas suffisant

Pour remédier à cela, certaines instances mettent en place des ensembles de lois avec pour objectif de protéger la confidentialité de leurs résidents. Nous retrouvons notamment les RGPD (*General Data Protection Regulation*) en Europe et le CCPA (*California Consumer Privacy Act*) en Californie. Néanmoins ces ensembles de lois ne sont pas suffisants. En effet, ils ne sont pas applicables au monde entier et surtout, ils ne préservent pas vraiment la confidentialité. Le récépissé 26 des GDPR autorise la conservation des données anonymisées si la condition *très subjective* suivante est respectée : les données ne permettent pas d'identifier la personne naturelle à l'aide de moyens raisonnables [PE16]. En pratique, cela revient à accepter que l'anonymisation des données est suffisante pour respecter la loi : les grands réseaux sociaux refusent de supprimer définitivement les messages des utilisateurs qui quittent la plateforme alors que de tels messages permettent *très* facilement de remonter à l'auteur. Pour donner un autre exemple plus précis : des chercheurs du MIT et de l'Université catholique de Louvain, ont montré, après avoir étudié les données de 1.5 millions de portables pendant 15 mois, que quatre points spatiaux relativement peu précis suffisent à identifier 95% des utilisateurs [Mon+13].



Encore pire, LATANYA SWEENEY a montré [Swe00] qu'en 1990 le ZIP-code, le genre (l'étude étant assez ancienne, il n'est pas clair si l'autrice parlait de genre ou de sexe) et la date de naissance suffisait à identifier 87% de la population américaine. Le lieu de naissance, le genre et la date de naissance permettent déjà d'identifier la moitié de la population alors que ces données sont couramment incluses dans les données anonymes !

1.1.3 L'appel à la *differential privacy*

L'anonymisation ne suffisant pas à réaliser des études statistiques de manière confidentielle, la *differential privacy* a été introduite de manière à quantifier la perte de confidentialité engendrée par une étude. Cette quantification permet d'étudier de manière précise de mécanismes et de fournir des réelles garanties mathématiques de confidentialité. L'introduction d'aléatoire permet de donner des réponses statistiques précises tout en assurant qu'il n'est pas possible de déduire la présence ou l'absence d'un individu du jeu de donné à partir de la réponse.

1.1.4 Contenu de ce rapport

Dans un premier temps nous donnerons une introduction aux fondements de la *differential privacy* et nous montrerons que la méthode naïve est inefficace ici. Ensuite nous présenterons la méthode que nous avons mis au points et nous l'étudierons en détail. Nous feront alors de même avec le meilleur algorithme actuel. Enfin, nous comparerons les deux méthodes.

1.2 Background essentiel sur la *differential privacy*

La *differential privacy* [Dwo+06] quantifie la perte de confidentialité subit par un individu en étant dans une base de donné.

Définition 1.2.0.1 : Jeu de données voisins

On dit que deux jeux de données x et y sont voisins et on note $d_{\text{Ham}}(x, y) \leq 1$ si ils 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 $\mathcal{M} : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T}$ est (ε, δ) -*differentially private* si pour tout $S \subset \mathcal{T}$ mesurable,

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

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

Intuitivement, plus une réponse est susceptible de varier due à la présence d'un individu dans un groupe, plus une réponse est susceptible de révéler des informations confidentielles. Nous définissons alors la sensibilité d'une requête pour quantifier ce phénomène.

Définition 1.2.0.3 : Sensibilité d'une requête

Soit \mathcal{X} un ensemble, (\mathcal{T}, d) un espace mesuré et $f : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T}$ une requête. On appelle **sensibilité de f** la grandeur Δf que l'on définit de la manière suivante :

$$\Delta f = \sup_{x, y \in \mathcal{X}^{(\mathbb{N})}} \{d(f(x), f(y)) \mid d_{\text{Ham}}(x, y) = 1\}$$

De manière informelle, la sensibilité d'une fonction exprime à quel point modifier une valeur du jeu de donné peu modifier la valeur de retour de la fonction. Le mécanisme de LAPLACE [DR14] est imaginé sur ce principe. Ce mécanisme est le mécanisme de *differential privacy* le plus simple et un des premiers présentés. Du bruit est ajouté aux réponses en fonction de la sensibilité des requêtes.

Définition 1.2.0.4 : Mécanisme de LAPLACE

Soit \mathcal{X} un ensemble de base, $\varepsilon \in \mathbb{R}_+^*$, $n \in \mathbb{N}$ et $f : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathbb{R}$ une requête. Notons Δf la sensibilité de f et Lap la fonction qui associe une variable aléatoire suivant la loi de LAPLACE¹ dont la paramètre est donné en argument. On appelle **mécanisme de LAPLACE** la fonction

$$\mathcal{M}_{f, \varepsilon} : \begin{cases} \mathcal{X}^{(\mathbb{N})} & \rightarrow & \mathbb{R} \\ x & \mapsto & f(x) + \text{Lap}\left(\frac{\Delta f}{\varepsilon}\right) \end{cases}$$

1. Il s'agit de la loi de densité $x \mapsto 1/(2b) \exp(-|x|/b)$ où b est le paramètre donné en argument.

Ce mécanisme est *vraiment* très simple à implémenter et tourne en temps constant. Les langages les plus utilisés dans le domaine tels que Python ou Julia ont la distribution de LAPLACE déjà intégrée. D'autres langages populaires tel que la C++ permettent néanmoins une implémentation tout aussi simple car cette distribution est facilement obtenable à partir d'une distribution exponentielle, bien plus commune. L'algorithme a néanmoins un grand défaut : si la requête est très sensible (c'est par exemple le cas si comme pour les quantiles elle peut subir un effet de pallié) le bruit ajouté est grand et nous perdons beaucoup en précision par rapport à ce que permettent d'autres mécanismes. Nous allons le voir un peu plus tard lorsque nous mètrons cet algorithme en difficulté face à l'estimation de déciles.

Théorème 1.2.0.1 :

Le mécanisme de LAPLACE est *differentially private*.

*Démonstration*² : Les notations utilisées sont les mêmes que celles de la définition du mécanisme. Soit $\mathcal{S} \subset \mathbb{R}$ et $x, y \in \mathcal{X}^{(\mathbb{N})}$ voisins. Notons que,

$$\mathbb{P}(\mathcal{M}_{f,\varepsilon}(x) \in \mathcal{S}) = \mathbb{P}\left(f(x) + \text{Lap}\left(\frac{\Delta f}{\varepsilon}\right) \in \mathcal{S}\right) = \int_{\mathcal{S}} \exp\left(-\frac{\varepsilon|t + f(x)|}{\Delta f}\right) dt.$$

Or, par définition de la sensibilité, pour tout t ,

$$|t + f(x)| \stackrel{\text{IT}}{\leq} |t + f(y)| + |f(x) - f(y)| \leq |t + f(y)| + \Delta f.$$

Ainsi,

$$\mathbb{P}(\mathcal{M}_{f,\varepsilon}(x) \in \mathcal{S}) \leq \int_{\mathcal{S}} \exp\left(-\frac{\varepsilon|t + f(y)|}{\Delta f} + \varepsilon \frac{\Delta f}{\Delta f}\right) dt = e^\varepsilon \int_{\mathcal{S}} \exp\left(-\frac{\varepsilon|t + f(y)|}{\Delta f}\right) dt.$$

Finalement,

$$\mathbb{P}(\mathcal{M}_{f,\varepsilon}(x) \in \mathcal{S}) \leq e^\varepsilon \mathbb{P}(\mathcal{M}_{f,\varepsilon}(y) \in \mathcal{S}).$$

Une variable de LAPLACE est d'espérance nulle. La linéarité de l'espérance permet de rapidement voir que l'espérance de l'algorithme précédent est $f(x)$ et donc qu'appliquer l'algorithme précédent suffisamment de fois permet de retrouver la vraie réponse avec une bonne précision. Nous remarquons alors que nous avons perdu de la confidentialité : appliquer n fois un algorithme ε -DP n'est pas ε -DP. Le théorème suivant permet de quantifier cela. Ce théorème est un résultat central dans le domaine. Nous en donnons ici une version fiable³ qui la version couramment utilisée.

Théorème 1.2.0.2 : Théorème de composition (simple)

Soit \mathcal{X} un ensemble de base, $n \in \mathbb{N}$ un nombre de mécanismes, $(\mathcal{T}_i)_{i \leq n}$ des ensembles d'arrivée et $(\mathcal{M}_i : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T}_i)_{i \leq n}$ des mécanismes mutuellement indépendants respectivement ε_i -*differentially private*.

L'exécution des n mécanismes est $\left(\sum_{i=1}^n \varepsilon_i\right)$ -*differentially private*.

Démonstration : Considérons $(\mathcal{S}_i)_{i \leq n} \subset \prod_{i=1}^n \mathcal{T}_i$ et $x, x' \in \mathcal{X}^{(\mathbb{N})}$ tel que $d_{\text{Ham}}(x, x') \leq 1$.

$$\begin{aligned} \mathbb{P}(\mathcal{M}_1(x) \in \mathcal{S}_1 \wedge \mathcal{M}_2(x) \in \mathcal{S}_2 \wedge \dots) & \stackrel{\text{indé.}}{=} \prod_{i=1}^n \mathbb{P}(\mathcal{M}_i(x) \in \mathcal{S}_i) \\ & \stackrel{\text{DP}}{\leq} \prod_{i=1}^n e^{\varepsilon_i} \mathbb{P}(\mathcal{M}_i(x') \in \mathcal{S}_i) \\ & \stackrel{\text{indé.}}{=} \exp\left(\sum_{i=1}^n \varepsilon_i\right) \mathbb{P}(\mathcal{M}_1(x') \in \mathcal{S}_1 \wedge \dots) \end{aligned}$$

2. La démonstration n'est pas la même que celle située dans le livre de référence, elle est personnelle.

3. Nous avons aussi étudié le théorème de composition avancé. Néanmoins il n'est utile que dans le cas de la (ε, δ) -DP, que nous avons évité durant ce stage. De plus, la démonstration est très longue et je ne me la suis pas suffisamment appropriée pour la reproduire.

1.3 L'échec de la méthode naïve

0	0
0	0
1	0
1	1
1	1

Se contenter d'utiliser un bruit de LAPLACE est tentant et très facile ! Nous l'avons fait pour le principe. L'implémentation est rapide à faire et très courte (alors qu'elle est réalisée en C++ !). Pourtant cela est **sans intérêt** ici ! En effet, disons que nous travaillons sur un ensemble dont nous savons qu'il est inclus dans $[0, 1]$. Le jeu de données ayant $2n + 1$ valeurs dont les n premières valeurs sont des 0 et le reste des 1 et le jeu de données dont $n + 1$ valeurs sont des 0 et le reste des 1 sont voisins. Pourtant l'un a 0 pour valeur médiane et l'autre 1. La sensibilité de la médiane est donc de 1. Notons alors que, en notant f la fonction médiane et en reprenant les notation précédentes,

$$|f(x) - \mathcal{M}_{f,\varepsilon}| = |\text{Lap}(1/\varepsilon)|.$$

Ainsi,

$$\mathbb{E}(|f(x) - \mathcal{M}_{f,\varepsilon}|) = \int_{\mathbb{R}} \mathbb{P}(|\text{Lap}(1/\varepsilon)| > t) dt = \int_{\mathbb{R}_+} e^{-\varepsilon t} dt = \frac{1}{\varepsilon}.$$

L'espérance de l'erreur commise en utilisant cette méthode vaut donc $1/\varepsilon$ alors que la médiane se situe dans $[0, 1]$ et que les valeurs de ε intéressantes sont petites et même, généralement plus petites que 1. **Cette méthode ne peut donc pas convenir**, il nous faut utiliser mieux.

2 Méthode des histogrammes

Au cours de cette section nous allons d'abord présenter un algorithme à la base de notre méthode. Ensuite nous allons présenter notre méthode ainsi que divers résultats théoriques et expérimentaux de précisions d'icelle.

2.1 AboveThreshold

Répondre à de nombreuses requête est coûteux en confidentialité (comme l'a montré le théorème de composition [2]). Utiliser à algorithme naïf tel que le mécanisme de LAPLACE [Dwo+06] 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** [DR14] permet cela tout en ne payant en confidentialité que pour les requêtes qui dépassent le seuil.

```

1  AboveThreshold(database, queries, threshold, epsilon){
2      Assert("les requêtes sont toutes de sensibilité 1");
3      result = 0;
4      noisyThreshold = threshold + Lap(2/epsilon);
5      for(querie in queries){
6          nu = Lap(4/epsilon);
7          if(querie(D) + nu > noisyThreshold)
8              return result;
9          else
10             ++result; /* Aucune requête n'a dépassé le seuil */
11     }
12     return -1;
13 }
```

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 légèrement adaptée de l'algorithme initialement décrit par DWORK et ROTH [DR14, page 57]. Icelui a du sens d'un point de vue informatique mais rend le formalisme mathématiques compliqué et nous n'utiliseront pas les légers avantages de leur version.

Théorème 2.1.0.1 :

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

Remarque : La démonstration est une réécriture de celle du livre de référence [DR14, page 57]. Une réécriture nous semblait nécessaire afin d'avoir une démonstration propre.

Démonstration :

Soit $D, D' \in \mathcal{X}^{(\mathbb{N})}$ tels que $d_{\text{Ham}}(D, D') \leq 1$, $\{f_i\}_i = Q \in (\mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T} \subset \mathbb{R})^{\mathbb{N}}$ un ensemble de requêtes de sensibilité 1, $T \in \mathbb{R}$ un seuil, et $\varepsilon > 0$. On pose A la variable aléatoire $\text{AboveThreshold}(D, Q, T, \text{epsilon})$ et A' la variable aléatoire $\text{AboveThreshold}(D', Q, T, \text{epsilon})$.

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

$$g_k = \max_{i < k} \{f_i(D) + \nu_i\} \quad \text{et} \quad g'_k = \max_{i < k} \{f_i(D') + \nu_i\}.$$

Ces grandeurs représente la valeur plus grande comparée au seuil bruité avant l'indice k dans le cas de l'exécution sur D et de l'exécution 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 $i \in \mathbb{N}$,

$$y_i = f_i(D) \quad \text{et} \quad y'_i = f_i(D').$$

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 $4/\varepsilon$,

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

On pose alors $\hat{t} = t + g_k - g'_k$ afin d'obtenir,

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

Il est alors temps de poser $\hat{\nu} = \nu + g_k - g'_k + y'_k - y_k$ et de remarquer que

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

Par définition de l_2 et l_4 nous avons donc

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

L'inégalité triangulaire assure alors que

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

Les requêtes étant de sensibilité 1, nous avons

$$\begin{cases} 2 & \geq |g_k - g'_k| + |y'_k - y_k| \geq |g_k - g'_k + y'_k - y_k| = |\hat{\nu} - \nu| \\ 1 & = |g_k - g'_k| = |\hat{t} - t| \end{cases}$$

La croissance de l'intégrale assure finalement que,

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

Nous avons donc montré que,

$$\mathbb{P}(A = k) \leq \exp(\varepsilon) \int_{\mathbb{R}} \int_{g'_k - T}^{y'_k + \nu - T} l_2(t) l_4(\nu) dt d\nu.$$

En faisant la même transformation qu'au début de cette démonstration, nous obtenons,

$$\mathbb{P}(A = k) \leq \exp(\varepsilon) \mathbb{P}(A' = k).$$

2.2 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 instantiation 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.

```

1  HistogramMethod(database, epsilon, a, b){
2      steps = 1.5*n/log(n);
3
4      /* composition theorem */
5      epsilon /= 9;
6
7      result = {};
8      for(d in {1 ... 9}){ /* which decile */
9          T = d*card(database)/10;
10         for(i in {1 ... steps}){
11             fi = x -> card({element in x | element < i*(b-a)/steps});
12             queries.push_back(fi);
13         }
14         T = d*card(database)/10;
15         result.push_back(AboveThreshold(database, queries, T, epsilon)
16                             *(b-a)/steps});
17     }
18     return result;
19 }
```

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 2.2.0.1 :

HistogramMethod est ε -differentially private.

Démonstration : Les requêtes envoyées 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 évaluer sa précision. Cela ne sera pas évident car la précision de l'algorithme dépend beaucoup du jeu de données d'entrée.

Lemme 2.2.0.1 : AboveThreshold est (α, β) -accurate

Pour tout $\beta \in]0, 1[$, tout $x \in \mathcal{X}^{(\mathbb{N})}$, tout $\{f_i\}_i = Q \in (\mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T})^{\mathbb{N}}$, tout $\varepsilon > 0$, tout $T \in \mathbb{R}$, en posant $\alpha = 8(\log(k) + \log(2/\beta))/\varepsilon$ et $k = \text{AboveThreshold}(x, Q, T, \text{epsilon})$, on a, en reprenant les notations de l'algorithme,

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

Remarque : Ce lemme est due à [DR14, page 61]. Nous reprenons aussi la démonstration ici car la démonstration originale ne nous semble pas assez claire.

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

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

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

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

Mutatis mutandis,

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

Ainsi,

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

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,

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

Nous remarquons alors que

$$\mathbb{P}\left(\max_{i \leq k} |\nu_i| + |T - \hat{T}| < \alpha\right) \geq 1 - \mathbb{P}\left(\max_{i \leq k} |\nu_i| \geq \frac{\alpha}{2}\right) - \mathbb{P}\left(|T - \hat{T}| \geq \frac{\alpha}{2}\right) \geq 1 - \frac{\beta}{2} - \frac{\beta}{2}.$$

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.

2.3 Analyse de complexité

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

2.4 Analyse de précision - le cas de la distribution uniforme standard

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 manières de penser ce qu'est la valeur attendue : elle pourrait être la valeur des déciles de l'échantillon d'entrée. 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'entrée 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 étaient 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. De plus, cela nous semble philosophiquement plus pertinent.

2.4.1 Analyse de précision : borne exacte

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

Définition 2.4.1.1 : Fonction Beta incomplète (régularisée)

On appelle respectivement fonction beta incomplète et fonction beta incomplète régularisée les fonctions

$$B : \begin{cases} [0, 1] \times (\mathbb{R}_+^*)^2 & \rightarrow \mathbb{R}_+ \\ (x, \alpha, \beta) & \mapsto \int_0^x t^{\alpha-1} (1-t)^{\beta-1} dt \end{cases} \quad \text{et} \quad I_\bullet : \begin{cases} (\mathbb{R}_+^*)^2 & \rightarrow \mathbb{R}_+ \\ (\alpha, \beta) & \mapsto \frac{B(\bullet, \alpha, \beta)}{B(1, \alpha, \beta)} \end{cases}.$$

Définition 2.4.1.2 : Loi beta

On appelle loi beta de paramètre $(\alpha, \beta) \in \mathbb{R}_+^*$ 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 2.4.1.3 : Statistique d'ordre

Soit X un échantillon statistique de cardinal $n \in \mathbb{N}$. Pour tout $k \in \llbracket 1, n \rrbracket$ on note $X_{(k)}$ et on appelle **statistique d'ordre** de rang k la k -ème plus petite valeur de l'échantillon.

Théorème 2.4.1.1 : 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 $k \in \llbracket 1, n \rrbracket$. La k -ème statistique d'ordre de X , $X_{(k)}$ est distribuée suivant la loi beta de paramètre $(k, n - k + 1)$.

Démonstration : disponible en annexe [A.1].

Lemme 2.4.1.1 : *Estimation de l'écart 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 $[0,1]$, $\gamma \in [0, d_i^l]$ et $\alpha \in [0, n/10]$. Notons $(d_i^l)_i$ les déciles de la loi. Pour tout $i \in \llbracket 1, 9 \rrbracket$

$$\mathbb{P}([X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]) \geq I_{d_i^l + \gamma}(in/10 + \alpha, n - in/10 - \alpha + 1) - I_{d_i^l - \gamma}(in/10 - \alpha, n - in/10 + \alpha + 1).$$

Démonstration : Notons que

$$\mathbb{P}([X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]) = \mathbb{P}(X_{(in/10-\alpha)} \geq d_i^l - \gamma \wedge X_{(in/10+\alpha)} \leq d_i^l + \gamma).$$

Ainsi,

$$\mathbb{P}([X_{(in/10-\alpha)}, X_{(in/10+\alpha)}] \subset [d_i^l - \gamma, d_i^l + \gamma]) \geq \mathbb{P}(X_{(in/10-\alpha)} \geq d_i^l - \gamma) + \mathbb{P}(X_{(in/10+\alpha)} \leq d_i^l + \gamma) - 1.$$

Or, le théorème précédent assure que

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

Ce que permet bien de directement obtenir le résultat.

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

Théorème 2.4.1.2 : *(α, β) -précision de `HistogramMethod` dans le cas uniforme standard*

Soit $\beta \in [0, 1]$ et X un ensemble de n (tel que ⁴ $8 \log(3n/(\beta \log n))/\varepsilon) \leq n/10$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur $[0,1]$. Soit $\gamma \in [0, 0.1]$ et $i \in \llbracket 1, 9 \rrbracket$. Notons $(d_i^l)_i$ les déciles de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, 0, 1)`, $\alpha = 8 \log(3n/(\beta \log n))/\varepsilon$ et $k = 1.5n/\log n$. On a

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

Démonstration : 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

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

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

4. Cette condition sera précisée plus tard. C'est une minoration de n .

au moins $1 - \beta$. Le lemme précédent permet alors de conclure.

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 2.4.1.3 : Précision moyenne de *HistogramMethod*

Soit $\beta \in [0, 1]$ et X un ensemble de n (tel que ⁵ $8 \log(3n/(\beta \log n))/\varepsilon) \leq n/10$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur $[0, 1]$. Soit $\gamma \in [0, 0.1]$ et $i \in [1, 9]$. Notons $(d_i^l)_i$ les déciles de la loi. Posons A la variable aléatoire $\text{HistogramMethod}(X, \text{epsilon}, 0, 1)$, $\alpha = 8 \log(3n/(\beta \log n))/\varepsilon$. On a

$$\begin{aligned} \mathbb{E}(|A_i - d_i^l|) &\leq \frac{2 \log n}{3n} + d_i^l \beta + \int_0^{d_i^l} \left(1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1)\right) dt \\ &\quad + \int_0^{d_i^l} I_{d_i^l-t}(in/10 - \alpha, n - in/10 + \alpha + 1) dt \\ &\quad + \left(1 - d_i^l - \frac{2 \log n}{3n}\right) \left(1 + \beta - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1)\right). \end{aligned}$$

Démonstration : On pose $k = 1.5n/\log n$ et

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

Le théorème précédent assure que

$$\forall t \in [0, 2d_i^l] \quad F\left(t + \frac{1}{k}\right) := \mathbb{P}\left(|A_i - d_i^l| \leq t + \frac{1}{k}\right)$$

Ainsi,

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

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

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

Notons que

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

De plus, 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,

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

5. Cette condition sera précisée plus tard. C'est une minoration de n .

Nous avons alors bien démontré le résultat.

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 été faites pour qu'une utilisation d'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 que le choix $\beta = 1/(\sqrt{n} \log n)$ n'est "pas trop" éloignée de cette borne inf. Nous disposons alors du corollaire suivant.

Corollaire 2.4.1.1 : (*im*) *Précision moyenne de HistogramMethod*

Soit $\beta \in [0, 1]$ et X un ensemble de n (tel que $8 \log(3n/(\beta \log n))/\varepsilon) \leq n/10$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur $[0, 1]$. Soit $\gamma \in [0, 0.1]$ et $i \in [1, 9]$. Notons $(d_i^l)_i$ les déciles de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, 0, 1)`, $\alpha = 8 \log(3n/(\beta \log n))/\varepsilon$. On a

$$\begin{aligned} \mathbb{E}(|A_i - d_i^l|) &\leq \frac{2 \log n}{3n} + \frac{d_i^l}{\sqrt{n} \log n} + \int_0^{d_i^l} \left(1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1)\right) dt \\ &\quad + \int_0^{d_i^l} I_{d_i^l-t}(in/10 - \alpha, n - in/10 + \alpha + 1) dt \\ &\quad + \left(1 - d_i^l - \frac{2 \log n}{3n}\right) \left(1 + \frac{1}{\sqrt{n} \log n} - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1)\right). \end{aligned}$$

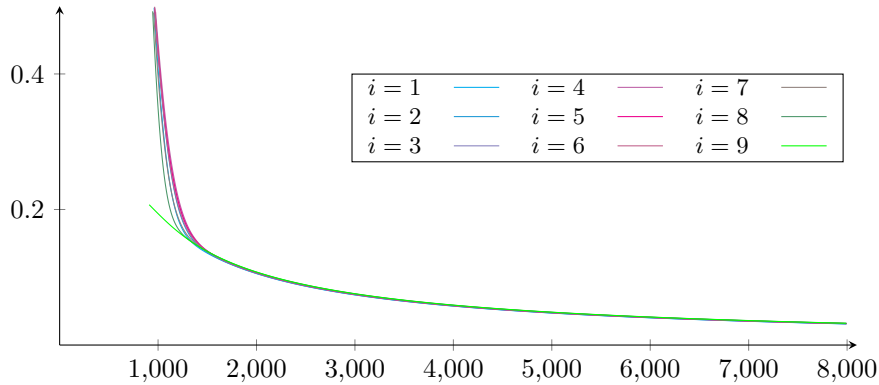


FIGURE 2 – Graphe de la borne du corollaire [1] en fonction de n avec $\varepsilon = 1$.

TODO justification borne

2.4.2 Analyse de précision : borne asymptotique

Le calcul d'une borne asymptotique sur l'espérance de la distance entre la sortie de l'algorithme et les déciles de la loi est obtenue à partir d'une majoration asymptotique du résultat du corollaire précédent [1]. Obtenir cette borne n'a pas été facile, il a fallu effectuer de nombreux essais avant de trouver une solution convenable : beaucoup de méthodes ne permettent pas une bonne simplification et fait alors obtenir une borne qui tend vers $+\infty$, une borne inutile ! Cette sous-sous-section présente le résultat que nous avons finalement réussi à obtenir.

Soit X un ensemble de n (tel que $8 \log(3n\sqrt{n})/\varepsilon) \leq n/20$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur $[0, 1]$. Soit $i \in [1, 9]$. Notons $(d_i)_i$ les décile de la loi. Posons

6. Cette condition sera précisée plus tard. C'est une minoration de n .

A la variable aléatoire `HistogramMethod(X, epsilon, 0, 1)` et $\alpha = 8 \log(3n\sqrt{n})/\varepsilon$.

Nous allons commencer par majorer le terme suivant :

$$\int_0^{d_i^l} \left(1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1)\right) dt.$$

Soit $t \in [0, d_i^l]$. Notons que

$$\begin{aligned} 1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1) &= I_{1-(d_i^l+t)}(n - (in/10 + \alpha - 1), (in/10 + \alpha - 1) + 1) \\ &\stackrel{\text{déf}}{=} I_{1-p}(n - k, k + 1). \end{aligned}$$

Or, si X suit une loi binomial de paramètres n, p , $I_{1-p}(n - k, k + 1) = \mathbb{P}(X \leq k)$. Nous souhaitons appliquer une application de l'inégalité d'HOEFFDING (démontré en annexe [A.2]). Pour cela il nous faut $k \leq np$. Il suffit pour cela que nous ayons $\alpha \leq nt$. Se restreindre aux $t \geq 12 \log(3^{2/3}n)/(n\varepsilon)$ permet donc d'appliquer l'inégalité. On obtient alors

$$1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1) \leq \exp \left(-2n \left(p - \frac{k}{n} \right)^2 \right).$$

Ainsi,

$$1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1) \leq \exp \left(-2 \frac{k^2}{n} \right) \exp(-2np^2 + 4pk).$$

De plus,

$$\int_0^{d_i^l} \exp(-2np^2 + 4pk) dt = \int_{d_i^l}^{2d_i^l} \exp(-2nt^2 + 4tk) dt \leq \int_{\mathbb{R}} \exp(-2nt^2 + 4tk) dt.$$

On a alors,

$$\int_0^{d_i^l} \exp(-2np^2 + 4pk) dt \leq \int_{\mathbb{R}} \exp \left(-2n \left(t - \frac{k}{n} \right)^2 + 2 \frac{k^2}{n} \right) dt = \frac{1}{\sqrt{2n}} \exp \left(2 \frac{k^2}{n} \right) \int_{\mathbb{R}} \exp(-t^2) dt.$$

Enfin,

$$\int_0^{d_i^l} \exp(-2np^2 + 4pk) dt \leq \sqrt{\frac{\pi}{2n}} \exp \left(2 \frac{k^2}{n} \right).$$

L'intégrande étant plus petite que 1 pour les valeurs de t que nous avons écartées, nous avons montré que,

$$\int_0^{d_i^l} \left(1 - I_{d_i^l+t}(in/10 + \alpha, n - in/10 - \alpha + 1)\right) dt \leq \frac{12}{n\varepsilon} \log(3^{2/3}n) + \sqrt{\frac{\pi}{2n}}. \quad (2)$$

Nous pouvons alors entamer la majoration du terme suivant

$$\int_0^{d_i^l} I_{d_i^l-t}(in/10 - \alpha, n - in/10 + \alpha + 1) dt$$

Soit $t \in [0, d_i^l]$. Notons que

$$\begin{aligned} I_{d_i^l-t}(in/10 - \alpha, n - in/10 + \alpha + 1) &\stackrel{\text{déf}}{=} I_p(k + 1, n - k) \\ &= \mathbb{P}(X > k) \end{aligned}$$

Où X suit une loi binomiale de paramètre (n, p) . Nous souhaitons une nouvelle fois appliquer l'inégalité d'HOEFFDING (démontré en annexe [A.2]). Pour cela, il nous faut $k \geq np$. Il suffit alors

d'avoir $nt \geq \alpha + 1$ donc d'avoir $t \geq 12 \log(3^{2/3}n)/(\varepsilon n) + 1/n$. En se restreignant à ces valeurs de t nous avons donc

$$\begin{aligned}\mathbb{P}(X > k) &\leq \exp \left(-2np^2 \left(\frac{k}{np} - 1 \right)^2 \right) \\ &= \exp \left(-2\frac{k^2}{n} + 4kp - 2np^2 \right)\end{aligned}$$

Ainsi,

$$\begin{aligned}\int_0^{d_i^l} I_p(k+1, n-k) dt &= \int_0^{d_i^l} I_t(k+1, n-k) dt \\ &\leq \exp \left(-2\frac{k^2}{n} \right) \int_{\mathbb{R}} \exp(-2np^2 + 4kp) dt \\ &= \exp \left(-2\frac{k^2}{n} \right) \int_{\mathbb{R}} \exp \left(-2n \left(p - \frac{k}{n} \right)^2 + \frac{2k^2}{n} \right) dt \\ &= \sqrt{\frac{\pi}{2n}}\end{aligned}$$

Enfin, nous avons démontré que,

$$\int_0^{d_i^l} I_{d_i^l-t}(in/10 - \alpha, n - in/10 + \alpha + 1) dt \leq \frac{12}{\varepsilon n} \log(3^{2/3}n) + \frac{1}{n} + \sqrt{\frac{\pi}{2n}} \quad (3)$$

Il ne reste alors plus qu'à majorer le terme suivant

$$1 - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1)$$

Ainsi,

$$\begin{aligned}1 - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1) &= I_{1-2d_i^l}(n - (in/10 + \alpha - 1), (in/10 + \alpha - 1) + 1) \\ &\stackrel{\text{d\'ef}}{=} I_{1-p}(n - k, k + 1)\end{aligned}$$

La condition initiale sur n assure que l'on a $k \leq np$. Ainsi, l'inégalité d'HOEFFDING (démontré en annexe [A.2]) assure que

$$1 - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1) \leq \exp \left(-2n \left(p - \frac{k}{n} \right)^2 \right)$$

Or, n est tel que $24 \log(3^{2/3}n)/(\varepsilon n) \leq n/20$ $k/n < 3/2d_i^l$. Ainsi,

$$1 - I_{2d_i^l}(in/10 + \alpha, n - in/10 - \alpha + 1) \leq \exp \left(-\frac{n(d_i^l)^2}{2} \right) \quad (4)$$

Théorème 2.4.2.1 : (im)Précision moyenne de *HistogramMethod*

Nous pouvons enfin réunir tous ces résultats intermédiaires et énoncer le théorème.

Soit X un ensemble de n (tel que $0 \leq 8 \log(3n\sqrt{n})/\varepsilon) \leq n/20$) variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme sur $[0,1]$. Soit $i \in \llbracket 1, 9 \rrbracket$ et $k \in \mathbb{N}$. Notons $(d_i)_i$ les décile de la loi. Posons A la variable aléatoire *HistogramMethod*(X , ε , 0, 1) et $\alpha = 8 \log(3n\sqrt{n})/\varepsilon$.

$$\begin{aligned}\mathbb{E}(|A_i - d_i^l|) &\leq \frac{2 \log n}{3n} + \frac{d_i^l}{\sqrt{n} \log n} + 2\sqrt{\frac{\pi}{2n}} + \left(1 - d_i^l - \frac{2 \log n}{3n} \right) \left(\frac{1}{\sqrt{n} \log n} + \exp \left(-\frac{n(d_i^l)^2}{2} \right) \right) \\ &\quad + \frac{24}{\varepsilon n} \log(3^{2/3}n) + \frac{1}{n}\end{aligned}$$

Corollaire 2.4.2.1 : *(im)Précision moyenne de HistogramMethod*

Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi uniforme standard. Soit $i \in \llbracket 1, 9 \rrbracket$. Notons $(d_i)_i$ les décile de la loi. Posons A la variable aléatoire `HistogramMethod(X, epsilon, 0, 1)`.

$$\mathbb{E}(|A_i - d_i^l|) = \mathcal{O}_n\left(\frac{1}{\sqrt{n}}\right)$$

2.4.3 Analyse de précision : résultats expérimentaux**2.5 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 2.5.0.1 : *Fonction d'erreur*

On appel fonction d'erreur la fonction suivant :

$$\text{erf} : \begin{cases} \mathbb{C} & \rightarrow & \mathbb{C} \\ z & \mapsto & \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) dt \end{cases}$$

Lemme 2.5.0.1 : *Déciles de $\mathcal{N}(0, 1)$.*

Les déciles de $\mathcal{N}(0, 1)$, notés $(d_i^l)_i$ sont

$$\forall i \in \llbracket 1, 9 \rrbracket \quad d_i^l = \sqrt{2} \text{erf}^{-1}(2 \times 0.1i - 1)$$

Démonstration : Soit $i \in \llbracket 1, 9 \rrbracket$. On note que

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

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 [C].

Lemme 2.5.0.2 : *Estimation de l'écart 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 $i \in \llbracket 1, 9 \rrbracket$

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

Avec

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

Lemme 2.5.0.3 :

Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in \llbracket 1, 9 \rrbracket$ 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 2.5.0.1 : (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 $\gamma \in [0, d_i^l]$, $i \in \llbracket 1, 9 \rrbracket$, $k \in \mathbb{N}$ 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 , ϵ , k , a , b).

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

Avec

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

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* [AD20]. Le mécanisme considère l'inverse du nombre de valeurs à modifier dans un ensemble de donnée pour passer à un autre ensemble de donnée sur lequel la requête a une autre valeur recherchée. Cela définit alors l'utilité d'une valeur pour instancier le mécanisme exponentiel [MT07].

Définition 3.1.0.1 : *Longueur*

Soit $x \in \mathcal{X}^{(\mathbb{N})}$, $f : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T}$ 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 $f : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T}$ et $\epsilon \in \mathbb{R}_+$. Pour une mesure μ sur \mathcal{T} , on définit le mécanisme aléatoire $M(x)$ par sa fonction de densité

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

Il n'y a qu'en $f(x)$ que $\text{len}_f(x, \cdot)$ est nulle. Ainsi le dénominateur pourrait être petit est donné une grande probabilité à des valeurs distantes de $f(x)$. On [MT07] introduit alors une version lisse du

mécanisme.

Définition 3.1.0.3 : Longueur lisse

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

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

Définition 3.1.0.4 : Mécanisme de sensibilité inverse ρ -lisse

Soit $f : \mathcal{X}^{(\mathbb{N})} \rightarrow \mathcal{T}$ 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

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

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

On note que

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

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

3.2.1 Le cas général

L'article présentant le mécanisme de sensibilité inverse [AD20] 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 $(d_i^l)_i$.

Dans un premier temps, voici un théorème permettant d'estimer la probabilité que la réponse de l'algorithme soit au moins à une distance donnée des déciles *empiriques*. Ce théorème est démontré en annexe [B] car la démonstration est longue et n'est pas celle qui nous intéresse ici. Nous avons fait le parti pris de nous concentrer sur l'écart avec les déciles de la loi. C'est donc ce que nous ferons ici. Nous allons donc démontrer ici le second théorème, plus simple, qui est celui qui a le plus de sens ici.

Théorème 3.2.1.1 : Ecart avec les déciles empiriques

Soit $\gamma \in \mathbb{R}_+^*$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ 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.

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

Le théorème suivant est une simplification du théorème précédent que nous avons réalisé.

Théorème 3.2.1.2 : Ecart avec les déciles théoriques

Soit $\gamma \in \mathbb{R}_+^*$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ 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 $(d_i^l)_i$ les déciles de la loi. Notons alors enfin M_{cont} le mécanisme de sensibilité inverse ρ -lisse.

$$\mathbb{P}(|M_{\text{cont},i} - d_i^l| > 2u + \rho) \leq \frac{R}{2\rho} \exp(-nup_{\min,i}\varepsilon/4) + \frac{2\gamma}{u} \exp\left(-\frac{1}{8}nup_{\min,i}\right)$$

Démonstration : Ce théorème donne une borne exponentielle sur la précision de l'algorithme. La 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 , $N_j \geq nup_{\min,i}/2$ ”.

$$\begin{aligned} \mathbb{P}(|M_{\text{cont},i} - d_i^l| > 2u + \rho) &= \mathbb{P}(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A) \mathbb{P}(A) \\ &\quad + \mathbb{P}(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid \bar{A}) \mathbb{P}(\bar{A}) \\ &\leq \mathbb{P}(|M_{\text{cont},i} - d_i^l| > 2u + \rho \mid A) + \mathbb{P}(\bar{A}) \end{aligned}$$

Nous savons que si l'événement A survient, pour tout t tel que $|t - d_i^l| > 2u$, au moins $nup_{\min,i}/2$ éléments séparent d_i^l 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 $|s - d_i^l| > 2u + \rho$, $\text{len}_f^\rho(x, s) \geq nup_{\min,i}/2$. Enfin, pour tout t tel que $|t - d_i^l| > 2u + \rho$,

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

Ainsi,

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

Finalement, il ne nous reste plus qu'à minorer $\mathbb{P}(A)$! Pour cela, pour tout $k \leq n - 1$ on pose $Z_k = \mathbb{1}_{x_k \in I_j}$ 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 borne de CHERNOFF (démontré en annexe [A.3]) assure enfin que

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

Enfin,

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

On obtient alors

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

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

3.3 Analyse de précision - le cas de la loi uniforme standard

Toujours dans une optique de comparaison entre ce mécanisme est celui que nous avons présenté, nous allons maintenant particulariser notre étude au cas particulier de la loi normale standard. Nous allons réécrire le résultat précédent dans ce cas particulier et calculer une majoration de l'espérance de l'erreur sur le calcul d'un décile, comme nous l'avons fait pour notre mécanisme. Pour commencer, énonçons le théorème précédent [1] dans notre cas particulier.

Théorème 3.3.0.1 : *Ecart avec les déciles théoriques*

Soit $u \in [0, 1/2]$, $\rho \in \mathbb{R}_+$ et $X \in [0, 1]^n$ dont les éléments sont obtenues à partir de la loi uniforme standard. On note $(d_i^l)_i$ les déciles de la loi. Notons alors enfin M_{cont} le mécanisme de sensibilité inverse ρ -lisse.

$$\mathbb{P}(|M_{\text{cont},i} - d_i^l| > 2u + \rho) \leq \frac{1}{2\rho} \exp(-nu\varepsilon/4) + \frac{4}{u} \exp\left(-\frac{1}{8}nu\right)$$

Nous pouvons alors enfin essayer de majorer $\mathbb{E}(|M_{\text{cont},i} - d_i^l|)$ à l'aide de ce résultat. Voici directement le grand théorème de cette section.

Théorème 3.3.0.2 :

Soit $\rho \in]0, 1 - 4/\sqrt{n}]$ et $X \in [0, 1]^n$ dont les éléments sont obtenues à partir de la loi uniforme standard. On note $(d_i^l)_i$ les déciles de la loi. Notons alors enfin M_{cont} le mécanisme de sensibilité inverse ρ -lisse.

$$\mathbb{E}(|M_{\text{cont},i} - d_i^l|) \leq \rho + \frac{4}{\sqrt{n}} + \frac{4}{n\varepsilon\rho} \exp\left(-\frac{\sqrt{n}\varepsilon}{2}\right) + \frac{16}{\sqrt{n}} \exp\left(-\frac{\sqrt{n}}{4}\right)$$

Démonstration : On pose

$$F : \begin{cases} \mathbb{R}_+ & \rightarrow & [0, 1] \\ t & \mapsto & \mathbb{P}(|M_{\text{cont},i} - d_i^l| \geq t) \end{cases}$$

Rappelons directement que

$$\mathbb{E}(|M_{\text{cont},i} - d_i^l|) = \int_0^{+\infty} F(t)dt = \int_0^1 F(t)dt$$

Nous pouvons alors découper notre intégrale de manière à diviser les calculs en fonction des termes prépondérants.

$$\mathbb{E}(|M_{\text{cont},i} - d_i^l|) = \int_0^1 F(t)dt = \int_0^{\rho+4/\sqrt{n}} F(t)dt + \int_{\rho+4/\sqrt{n}}^1 F(t)dt$$

Ainsi

$$\mathbb{E}(|M_{\text{cont},i} - d_i^l|) \leq \rho + \frac{4}{\sqrt{n}} + \int_{\rho+4/\sqrt{n}}^1 F(t)dt \leq \rho + \frac{4}{\sqrt{n}} + \int_{2/\sqrt{n}}^{1/2-\rho} F(2u + \rho)du$$

Nous avons donc montré que

$$\begin{aligned}\mathbb{E}(|M_{\text{cont},i} - d_i^l|) &\leq \rho + \frac{4}{\sqrt{n}} + \int_{2/\sqrt{n}}^{1/2} \left(\frac{1}{2\rho} \exp(-nu\varepsilon/4) + \frac{4}{u} \exp\left(-\frac{1}{8}nu\right) \right) du \\ &\leq \rho + \frac{4}{\sqrt{n}} + \left(\frac{4}{n\varepsilon\rho} \exp\left(-\frac{\sqrt{n}\varepsilon}{2}\right) + \int_{2/\sqrt{n}}^{1/2} 2\sqrt{n} \exp\left(-\frac{1}{8}nu\right) du \right)\end{aligned}$$

Enfin,

$$\mathbb{E}(|M_{\text{cont},i} - d_i^l|) \leq \rho + \frac{4}{\sqrt{n}} + \frac{4}{n\varepsilon\rho} \exp\left(-\frac{\sqrt{n}\varepsilon}{2}\right) + \frac{16}{\sqrt{n}} \exp\left(-\frac{\sqrt{n}}{4}\right)$$

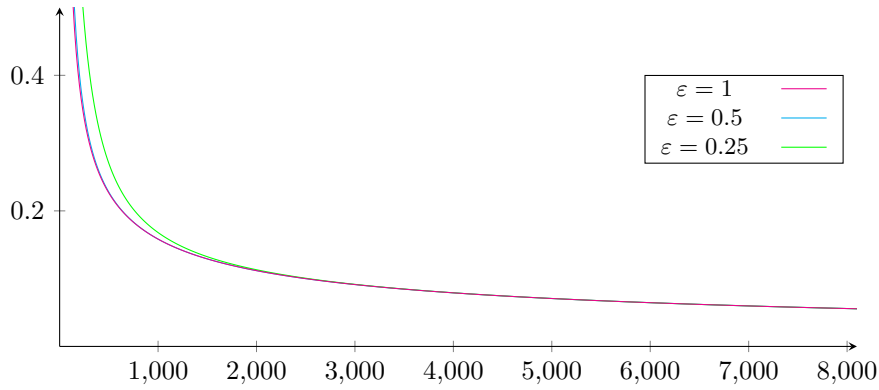


FIGURE 3 – Graphe de la borne obtenue [1] en fonction de n pour $\rho = 1/\sqrt{n}$.

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 Comparaison des bornes obtenues

4.2 Résultats expérimentaux

4.2.1 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 $n \in \llbracket 100, 5000 \rrbracket$ nous avons lancé les deux algorithmes sur 50 ensembles de données indépendants et identiquement distribué suivant $\mathcal{U}(0, 1)$.

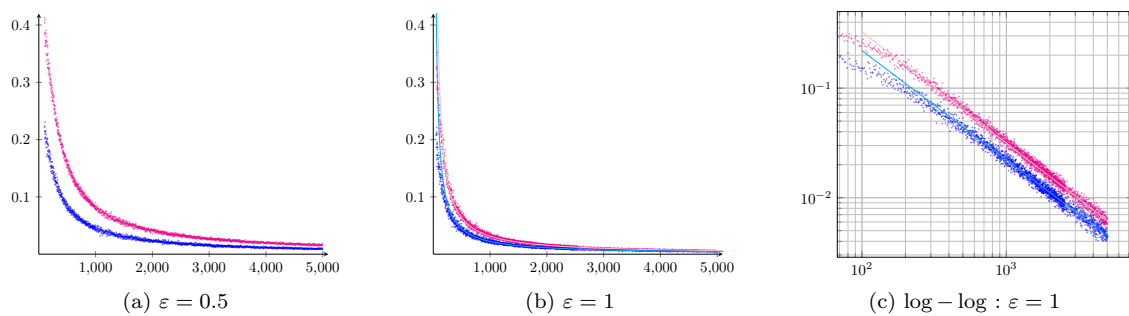


FIGURE 4 – Écart-quadratique moyen sur le calcul des déciles en fonction de n (la taille de l'échantillon). La méthode des histogrammes est en **magenta** et le mécanisme de sensibilité inverse est en **bleu**.

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.

4.2.2 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.

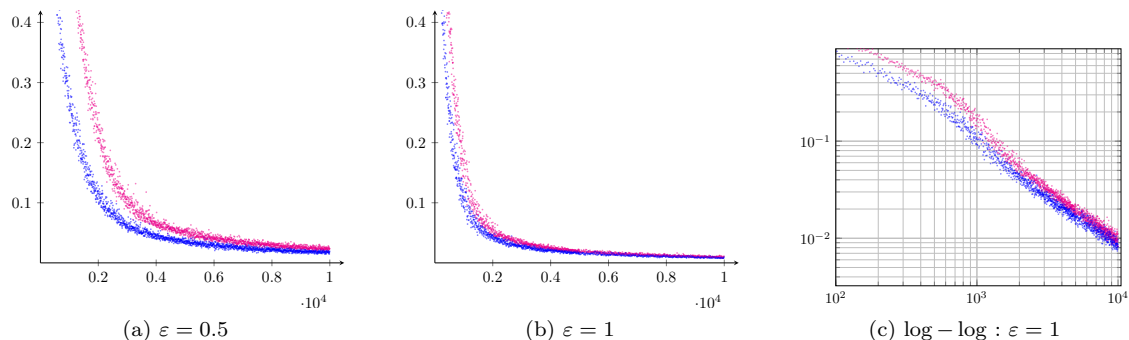
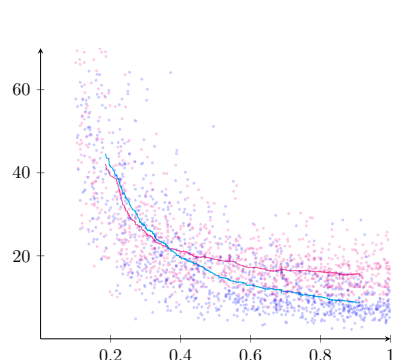


FIGURE 5 – Écart-quadratique moyen sur le calcul des déciles en fonction de n (la taille de l'échantillon). La méthode des histogrammes est en **magenta** et le mécanisme de sensibilité inverse est en **bleu**.

4.2.3 Des données réelles



Nous avons décidé de comparer les résultats sur une base de données réelle [NY21] afin de comparer les deux méthodes.

comment les méthodes se com-
portent vraiment en pratique. Pour
cela nous avons utilisé les sa-
lares annuels des agents du
publique de l'état de New-
York (États-Unis). Ces don-
nées ont directement été pu-
bliques par une agence gouver-
nementale, ce qui assure une
certaine fiabilité. Nous tra-
vaillons alors sur un ensemble
de près de 400 000 salaires an-
nuels.

Les courbes sur le graphe 6 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 6 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.

Références

- ASI, Hilal et John C. DUCHI. “Near Instance-Optimality in Differential Privacy”. In : *ArXiv* abs/2005.10630 (mai 2020). URL : <https://arxiv.org/pdf/2005.10630.pdf>.
- AUXIER, Brooke et al. *Americans and Privacy : Concerned, Confused and Feeling Lack of Control Over Their Personal Information*. 15 nov. 2019. URL : https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/11/Pew-Research-Center_PI_2019.11.15_Privacy_FINAL.pdf (visité le 20/07/2022).
- DWORK, Cynthia et Aaron ROTH. “The Algorithmic Foundations of Differential Privacy”. In : *Foundations and Trends in Theoretical Computer Science* 9 (août 2014), p. 211-407. URL : <https://www.microsoft.com/en-us/research/publication/algorithmic-foundations-differential-privacy/>.
- DWORK, Cynthia et al. “Calibrating Noise to Sensitivity in Private Data Analysis”. In : *Theory of Cryptography*. Sous la dir. de Shai HALEVI et Tal RABIN. Berlin, Heidelberg : Springer Berlin Heidelberg, 2006, p. 265-284. ISBN : 978-3-540-32732-5.
- MCSherry, Frank et Kunal TALWAR. “Mechanism Design via Differential Privacy”. In : *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, oct. 2007. URL : <https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/>.
- MEEHAN, Mary. *Data Privacy Will Be The Most Important Issue In The Next Decade*. 26 nov. 2019. URL : <https://www.forbes.com/sites/marymeehan/2019/11/26/data-privacy-will-be-the-most-important-issue-in-the-next-decade/?sh=430b3e591882> (visité le 20/07/2022).
- MONTJOYE, Yves-Alexandre de et al. “Unique in the Crowd : The privacy bounds of human mobility”. In : *Nature* 3 (mars 2013). URL : <https://doi.org/10.1038/srep01376>.
- NY, Open Data. *Salary Information for Local Authorities*. Authorities Budget Office. 13 déc. 2021. URL : <https://data.ny.gov/Transparency/Salary-Information-for-Local-Authorities/fx93-cifz> (visité le 08/07/2022).
- PARLIAMENT, European et Council of the EUROPEAN UNION. *Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive)*. 14 avr. 2016. URL : <https://gdpr-info.eu/recitals/no-26/> (visité le 20/07/2022).
- SWEENEY, Latanya. “Simple Demographics Often Identify People Uniquely”. In : (jan. 2000). DOI : 10.1184/R1/6625769.v1. URL : <https://dataprivacylab.org/projects/identifiability/paper1.pdf> (visité le 20/07/2022).

A Démonstration de théorèmes utiles

A.1 Loi des statistiques d'ordre

TODO

A.2 Inégalité d'HOEFFDING

TODO

A.3 Bornes multiplicatives de sc Chernoff

TODO

A.4 Déciles de la loi normale centrée réduite

TODO

B Le mécanisme de sensibilité inverse - probabilité de précision sur les déciles empiriques

Théorème B.0.0.1 :

Soit $\gamma \in \mathbb{R}_+^*$, $u \in [0, \gamma/4]$, $\rho \in \mathbb{R}_+$ 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.

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

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 " $|d_i^l - d_i| \geq \gamma/2$ ".

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

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

$$|t - d_i| > 2u + \rho,$$

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

Ainsi,

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

Nous allons maintenant calculer la probabilité de l'événement \overline{B}_i . Pour cela, on pose $\alpha = \gamma/2$, pour tout $j \in \llbracket 0, n-1 \rrbracket$ on pose $C_j^i = \mathbf{1}_{x_i > d_i^l + \alpha}$ 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

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

Or, si $d_i > d_i^l$, $C^i \geq in/10$. Ainsi, en utilisant une borne de CHERNOFF (démontré en annexe [A.3])

(C^i est d'espérance $\hat{p}n$ et les $(C_j^i)_j$ sont indépendantes),

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

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

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

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

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

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 (démontré en annexe [A.3]) assure enfin que

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

Enfin,

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

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 !

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

Lemme C.0.0.1 : *Estimation de l'écart 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 $i \in \llbracket 1, 9 \rrbracket$

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

Avec

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

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 $j \in \llbracket 0, n-1 \rrbracket$ 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 (démontré en annexe [A.3]) assure alors que

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

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

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

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

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

Finalement,

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

Lemme C.0.0.2 :

Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in \llbracket 1, 9 \rrbracket$ 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 $\gamma \in [0, d_i^l]$, $i \in \llbracket 1, 9 \rrbracket$ 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 $j \in \llbracket 0, n-1 \rrbracket$ 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 (démontré en annexe [A.3]) assure alors que

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

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,

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

Théorème C.0.0.1 : (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 $\gamma \in [0, d_i^l]$, $i \in \llbracket 1, 9 \rrbracket$, $k \in \mathbb{N}$ 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 , ϵ , k , a , b).

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

Avec

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

Démonstration : Soit X un ensemble de n variables aléatoires $(X_i)_i$ indépendantes et suivant toutes la loi normale centrée réduite. Soit $\gamma \in [0, d_i^l]$, $i \in \llbracket 1, 9 \rrbracket$, $k \in \mathbb{N}$ 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

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

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

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

Les lemmes précédent assurent alors que

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

Avec

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